Where is Cognition?

Towards an Embodied, Situated, and Distributed Interactionist Theory of Cognitive Activity

A thesis

submitted in partial fulfilment

of the requirements for the Degree

of

Doctor of Philosophy in Psychology

in the

University of Canterbury

by

Stephen Hill

University of Canterbury

2000

Table of Contents

Acknowledgements iv

Abstract v

1. Cognition: A Concept in Transition 1

Introduction 1

Characterisations of Cognition 5

The Problem of Cognitive Order 9

Cognition as Time-Space Distancing 12

Conclusion 22

2. Cognitivism and its Problems 23

Introduction 23

Using Representation and Computation to Explain Time-Space Distancing 26

The Sense-Model-Plan-Act Schema and Formal Task Description 60

Conclusion: Cognitivism as a Methodological Framework 75

3. Embodied, Situated, and Distributed Interactionist Influences 80

Introduction: The Emergence of ESD interactionism 80

Influential Fields 82

Summary 118

4. Embodied, Situated, and Distributed Interactionist Principles 120

The Main Principles of the Interactionist Approach 120

Explaining Wild Cognition 154

5. Explaining Wild Cognition with Dynamical Systems Theory 158

Introduction 158

The Promise of Dynamical Systems Theory 158

Does the Dynamical Approach Provide a Genuine Explanatory Framework? 186

What does Dynamical Systems Theory do for us? 190

6. What Kind of Dynamical System is a Cognizer? 192

Introduction 192

Autonomous Systems and the Dynamics of Being Alive 193

The Role of the Nervous System in Cognitive Activity 208

The Autonomous Systems View in Summary 227

7. Connectionism as a Model of the Embodied Nervous System 229

Introduction: Connectionism meets ESD Interactionism 229

The Connectionist Profile 230

Some Problems with Traditional Connectionism 234

Embodying and Situating Connectionism 239

The Limits of Connectionism 248

8. From Basic to Advanced Cognition 257

Introduction 257

Advanced Cognition and Imagery 266

Learning to Control Imagery 273

Conclusion 290

9. Where is Cognition? 294

The Lessons of Interactionism 295

Studying Cognition Interactionist Style 299

References 303

Acknowledgements

Many people provided input and support in the production of this thesis and I would like to thank them all. My supervisor Brian Haig provided support, encouragement, and guidance throughout. In him I was lucky enough to find a critical, multidisciplinary, and multitalented psychologist who thinks that theoretical work is a worthwhile endeavour. I have benefited from the ideas and musings of Canterbury academics and visitors to our department. Particular thanks go to Dean Owen and Lucy Johnston for inviting me to be one of the 'gang of four' in the social-ecological psychology group (or is that the ecological-social group?). Several people sent me manuscripts, papers, and ideas from various places around the world – thanks to John Flach, Alan Costall, Fred Keijzer, and Markus Latzina. A number of non-psychology university staff have been good enough to give me advice and let me sit in on their courses and seminars. They include Andrew Carstairs McCarthy, Robyn Schafer (linguistics), and Derek Browne (philosophy). Thanks to Mary-Anne (TAC2) and Murray Simmonds for making it easy to do my Ph.D. when I should have been worrying about coordinating the first year lab course. To all my friends and family, near and far, who have supported and encouraged me during the writing of this thesis a big thank you. Special thanks to Joan for all the coffee and conversation and to Julie for everything.

Abstract

In recent years researchers from a variety of cognitive science disciplines have begun to challenge some of the core assumptions of the dominant theoretical framework of cognitivism including the representation-computational view of cognition, the sense-model-plan-act understanding of cognitive architecture, and the use of a formal task description strategy for investigating the organisation of internal mental processes. Challenges to these assumptions are illustrated using empirical findings and theoretical arguments from the fields such as situated robotics, dynamical systems approaches to cognition, situated action and distributed cognition research, and sociohistorical studies of cognitive development. Several shared themes are extracted from the findings in these research programmes including: a focus on agent-environment systems as the primary unit of analysis; an attention to agent-environment interaction dynamics; a vision of the cognizer's internal mechanisms as essentially reactive and decentralised in nature; and a tendency for mutual definitions of agent, environment, and activity. It is argued that, taken together, these themes signal the emergence of a new approach to cognition called *embodied, situated, and distributed interactionism*. This interactionist alternative has many resonances with the dynamical systems approach to cognition. However, this approach does not provide a theory of the implementing substrate sufficient for an interactionist theoretical framework. It is suggested that such a theory can be found in a view of animals as autonomous systems coupled with a portrayal of the nervous system as a regulatory, coordinative, and integrative bodily subsystem. Although a number of recent simulations show connectionism's promise as a computational technique in simulating the role of the nervous system from an interactionist perspective, this embodied connectionist framework does not lend itself to understanding the advanced 'representation hungry' cognition we witness in much human behaviour. It is argued that this problem can be solved by understanding advanced cognition as the re-use of basic perception-action skills and structures that this feat is enabled by a general education within a social symbol-using environment.

1. Cognition: A Concept in Transition

Cognition may not be what we think it is.

Ed Hutchins (according to Elman, 1995, p. 195)

Introduction

The majority of psychologists and other cognitive scientists have traditionally assumed that cognition is what the mind does, and that the term *mind* is just an old-fashioned way of talking about the functioning of the brain. So the major focus for cognitive explanation has been on what goes on inside agents' heads. Ask a traditionalist where we might find cognition and you will receive a response something like this:

> The fact that cognition is something the brain does is so obvious it seems barely worth stating. (Kosslyn, 1986, p. 231).

I will refer to this dominant theoretical framework as *cognitivism*. Cognitivism underpins much of the theoretical work done in cognitive psychology and the other cognitive science disciplines such as artificial intelligence, linguistics, philosophy of mind, and neurobiology. And insofar as cognitive psychology has become the default foundation of nearly all modern psychology, cognitivism can be viewed as forming the foundation of most contemporary psychological theory. The cognitivist framework, as I shall characterise it, consists of three interlocking assumptions. First, the cognizer is assumed to be an *information-processor* who extracts, stores, and manipulates information about things in the environment. Second, information is considered to be realised by *representations* which are transformed and manipulated by *computations*. Finally, from a methodological perspective these representational-computational information-processing sequences and operations are arrived at via a process of *formal task description*. Roughly speaking, this involves working out what algorithms are logically necessary for carrying out various kinds of cognitive tasks and attempting to map these procedures on to neurocognitive structures and processes. I will have more to say about cognitivism in the next chapter. For now it is important to note that these three assumptions underpin cognitivism's commitment to the idea that cognition goes on in one's head.

Recently attitudes have begun to change. A growing number of researchers from various cognitive science-related disciplines have argued that many cognitive activities and behaviours are not best understood primarily in terms of the processes going on inside agents' heads (i.e., in their central nervous systems or brains). Instead, they argue that we must understand behaviour as arising from the continuous interplay of the internal resources (brains, bodily systems) and external resources (objects in the environment, other agents, and so on). Claims like the following are beginning to appear in the literature:

Minds are not purely internal things; they are, in part, worldly in character. That is, minds are hybrid entities, made up in part of what is going on inside the skin of creatures who have them, but also made up in part of what is going on in the environment of those creatures. (Rowlands, 1999, p. 29)

Enter the new wave in cognitive science[1]. It does not currently have a commonly agreed upon name, probably because it is really still a loose amalgamation of ideas, techniques, and disciplines. I will call this new approach *embodied, situated, and distributed interactionism (ESD interactionism* for short. Others call it (or something like it), *situated cognition* (Clancey, 1997), *embodied cognition* (Clark, 1997*), embedded cognition* (Haugeland, 1995), *enactivism* (Varela, Thompson & Rosch, 1991), *environmentalism* (Rowlands, 1999), *interactivist-constructionism* (Christensen & Hooker, 2000, in press) and *behavioural systems theory* (Keijzer, 1997). There is a feeling amongst some that ESD interactionism may revolutionise cognitive science.

We contend ... that a breakthrough is occurring within cognitive science that will prove to be a more dramatic turning-point in the history of western ideas than the 'cognitive revolution' ever was. We are witnessing a shift from a linguistic and formalistic conception of mind and intelligence, to an approach in which mind is conceived as sensitivity and adaptivity to the environment. Cognitive science is presently breaking with the fundamental and exclusive definition of the human being as the '*animal rationale*' (the thinking or intellectual animal). Mind is no longer regarded as a disembodied rational entity, immersed in its own thoughts, but is seen as being active in the world. Mind is seen no longer as the manipulator of static, pre-encoded representations in and for itself, but as taking part in dynamic and adaptive relations with the outside world. Mind is not a network of self-propelling propositions, but a way of constructing meanings in the world. (Bem & Keijzer, 1996, pp. 449-450)

More specifically interactionism departs from cognitivism in the following ways:

1. Interactionists are generally sceptical about the usefulness of the notion of representation. Some interactionists are anti-representationalists (they actively argue against representation), others are non-representationalists (they simply use other theoretical concepts), while others are agnostic or endorse some highly-modified sense of representation. The key point of agreement amongst ESD-oriented researchers is the rejection of the notion of representations as models of the world and plans for action. I will refer to anti- and non-representationalists *as ESD interactionists* and those who continue to endorse the use of some form of representation as *ESD representationalists.*
2. Interactionists generally reject the sense-model-plan-act schema as a template for describing the way cognizers work. Many ESD researchers argue that the kind of analysis used to produce these information-processing stories is inherently flawed. Instead the functional architecture of agents is often understood as being made up of a wired together collection of parallel-operating behaviour layers (Brooks, 1991a,

[1] Albeit one with old roots in Dewey and others (see chapter 3).

1991b). Each layer carries out the internal operations necessary for generating a basic kind of behavioural activity. The structure and function of such layers cannot be discovered by employing formal task description. Instead interactionists suggest that we must engage in systematic behavioural analysis and real-world modelling (Hendriks-Jansen, 1996).

3. The primary focus of much interactionist research is on basic cognitive and behavioural phenomena. Higher-cognition is understood as deeply grounded in basic, skilful, sensorimotor ability. This effectively reverses traditional cognitivist thinking which models 'knowing how' as a low-level, unconscious version of inference and reasoning (see Bechtel & Abrahamsen [1991]). By contrast, cognitivism focuses upon high level human cognitive phenomena such as speech, complex problem-solving, and deductive reasoning. The archetypal subject of cognitivism is the normal, educated, and typically literate, human adult (Von Eckardt, 1993). Although these phenomena constitute legitimate areas of study for cognitive science the methodological and explanatory frameworks in which these phenomena are embedded have been carried over to studies of lower level cognitive phenomena such as automatic and unconscious processes, practical skills, and the activities of 'non-language users' such as infants, non-humans, and our hypothesised hominid ancestors. Whether or not this importation of concepts deriving from high-level human cognitive capacities is a legitimate explanatory move is the focus of much current controversy.

4. Interactionists reject the equation of 'the mind' with the brain. The brain is viewed primarily as a coordinative and regulative system that is coupled with the rest of the body and, via that, with the environment. Brain, body, and world interact in interesting ways and cognitive activity emerges as result. Cognition, if it is anywhere, is to be found embedded in the 'organism-in-the-environment'.

These ideas form the core of this thesis. The main aim of this work is to unravel the complex issues at the heart of psychology's present approach to the study of cognition in the hope of making better sense of the relevant phenomena. It appears that the current cognitivist framework is coming to the end of its useful life and is already in the process of evolving into something quite different. The implications for research, and scientific and social policy based on that research, are profound and far-reaching. However this work is neither an attempt at building a comprehensive theoretical alternative to present orthodoxy in psychology and cognitive science, nor a procedural scheme suggesting better ways for doing research, although it touches on both of these points. Rather, it is an attempt to encourage psychologists to think differently about the concept of cognition and how and why we study and use it.

An Outline of the Structure of this Work

The bulk of this thesis is an attempt to sketch out the commitments of an interactionist framework for understanding cognition and to contrast these with those of cognitivism. In particular, the aim here is to investigate the empirical and theoretical research implications that are posed when the concept of cognition is understood in an interactionist manner. My goals are relatively modest. In the final part of this chapter I begin by arguing that cognition can be usefully understood in terms of the 'time-space distancing skills and capacities' animals have for taking account of and responding to aspects of local and distant (including hypothetical and imaginary) environments. In chapter 2 I examine the cognitivist framework and describe several problems that it faces. In particular I focus on the ways in which the representational metaphor may lead us *methodologically* astray by guiding us toward individualistic, in-the-head explanations and explorations of cognitive activity. It is possible to think about the relationship between the embodied agent and the world in a non-representational manner by recasting the representational notion of 'constructing internal content' as the possession of an ability to 'be sensitive to'. This rather vague idea seems to be operative in many interactionist accounts of cognition. Chapter 3 outlines several fields of recent cognitive scientific research that challenge many of the basic assumptions of cognitivism. These research areas include situated robotics, active vision research, dynamical approaches to cognition, distributed cognition, and sociohistorical approaches to development. In chapter 4 I distil several key themes from the insights of these fields including the concepts of the agent-environment system, perception-action cycles, reactive-decentralised architectures, and interactive emergence. I go on in chapter 5 to argue that dynamical systems theory provides a promising framework for making sense of cognition from an ESD perspective but that, in addition, an interactionist cognitive psychology requires a 'theory of the implementing substrate'. Chapter 6 suggests that such a theory can be found in an 'autonomy view' of animals that portrays the nervous system as a regulatory, coordinative, and integrative bodily subsystem. Chapter 7 describes how several recent simulations show connectionism's promise as a computational technique in simulating the role of the nervous system from an interactionist perspective. Unfortunately, as it stands, such a framework does not seem to lend itself to understanding the advanced 'representation hungry' cognition we witness in much human behaviour. So in chapter 8 I speculate that much of what we call advanced cognition involves a 're-use' of basic perception-action skills and structures. Furthermore, this reuse is the outcome of a general education within a symbol-using environment. One potentially fruitful way of unpacking this hypothesis is to suggest that all advanced cognizing is a form of 'mental imagery use' and that mental imagery is a 'socially guided' use of perception-action skills and structures.

In sum this work attempts to tie down some basic themes that will facilitate understanding cognitive phenomena from an interactionist perspective in the hope that new ways of theorising and researching may be encouraged. This is no easy task, as the modern view of cognition held by many psychologists and cognitive scientists is deeply intertwined with the cognitivist perspective. In order to alleviate this problem, the rest of this chapter is dedicated to unpacking what we mean when we talk of cognitive phenomena.

Characterisations of Cognition

The term cognition is used in a wide variety of ways by psychologists and other cognitive scientists. It seems to be most commonly understood by psychologists as a high-brow synonym for *thought* or *thinking*[2]. Thoughts then become *cognitions*. From this perspective cognition is just 'psychology-speak' for an aspect of the everyday (Western) understanding of mind. More specifically, cognition is frequently considered to feature in the age-old mental triad of cognition, emotion, and volition (conation). This distinction between rational thought, affect, and motivation to act is quite popular within other branches of psychology. In social psychology, for instance, attitudes are sometimes understood in terms of an ABC (affect, behaviour, cognition) model (Myers, 1993). However, it is clear that the introduction of the term *cognition* by cognitive revolutionaries such as Jerome Bruner and George Miller was not intended to be understood so narrowly. In an interview with Bernard Baars (1986), Miller notes that:

> I think they [the early cognitive psychologists] were just reaching back for common sense. In using the word 'cognition' we were setting ourselves off from behaviorism. We wanted something *mental* - but "mental psychology" seemed terribly redundant. "Commonsense psychology" would have suggested some sort of anthropological investigation, and "folk psychology" would have suggested Wundt's social psychology. What word do you use to label this set of views? We chose 'cognitive'.
>
> But "cognition" was meant in a very broad sense. When Jerry Bruner and I started the Center for Cognitive Studies at Harvard, did we mean to exclude anything that a computer can't do? Emotion, will, motivation? No, of course not. (p. 210).

So, for Miller at least, it is clear that the use of the term *cognition* served as more of a signpost of the break with behaviourism than the careful formulation of a new psychological phenomenon. Such a reading is implicit in the critical comments of a number of psychologists working in the transition period between behaviourism and cognitivism. Commenting on psychologists' early enthusiasm for the notion of cognition William Battig noted:

---

[2] This is admittedly only a personal observation garnered from going to many seminars, lectures, and presentations in academic environments and from reading academic texts and papers. Psychologists seem to use *cognition* in this manner more often than philosophers do, but I think that the term is generally adopted by many within the cognitive sciences.

What makes the current high popularity of cognition especially astounding is that even its most dedicated advocates seem unable to provide us with a clear or consistent definition of exactly what is meant by or encompassed under the cognitive label, or how it is to be distinguished from the allegedly noncognitive character of whatever is (or was) not described as cognitive psychology. (Battig, 1975, p. 195 quoted in Knapp, 1986, p. 29)

The popularity of the term cognition at the expense of a clear understanding of the concept itself led Ernest Hilgard to the only slightly cynical observation that cognition might be understood as "something you put in front of a book title that wouldn't have been there a couple of years ago" (in Fisher, 1983, p. 18 quoted in Knapp, 1986, p. 30). However, despite these critical observations *cognition* has become a popular word and has infiltrated most areas of modern psychology. Kessen (1981) goes so far as to say that psychology "has been redefined as the study of cognition." (p. 181):

Friendship has become social cognition, affect is seen as a form of problem-solving, newborn perception is subsumed under a set of transforming rules, and psychoanalysis is reread as a variant of information processing. Cognition, the feeble infant of the late Fifties and early Sixties, has become an insatiable giant. (p. 181).

Cognition has become such a ubiquitous term that many psychologists write as if its meaning is so obvious that the audience does not need to be told. More deviously, some researchers actually claim that the lack of a thorough characterisation of the central concepts of a discipline, such as cognition, should be expected because they "often prove to be the most difficult to analyze explicitly." (Palmer & Kimchi, 1986, p. 43)

In contrast to the rather loose ways in which psychologists use the term *cognition,* the philosophers of mind of the 1950s and 1960s, such as Fodor, Putnam, Armstrong, Sellars, and Chisholm, intended the adjective *cognitive* to be restricted to those kinds of mental states, such as beliefs and desires (propositional attitudes), that could be truth-evaluated, and not mental states such as consciousness and qualia (subjective experiences) (Green, 1996). For the philosophical psychologists (and the philosophically sophisticated Chomsky) cognition was not a synonym for *mental* but rather for that part of the mental that exhibits evaluable consequences in the external world. However, like Miller and Bruner, they wanted to construct a scientifically respectable alternative to behaviourism and wanted to use the concept of cognition as a central feature. Green (1996) argues that "the strict use of 'cognitive' [by the philosophers] was intended to keep the behaviourist criticisms of early-century mentalism at bay, [but that] its broad use in psychology returned us to precisely these criticisms." (p. 37-38). He goes on to argue that:

cognitivism was an answer to the problem: how can we introduce (at least part) of the mental back into scientific psychology while not falling prey to the criticisms that brought down the mentalism of old and led us to behaviourism? ... The answer that cognitivism has provided ... is that as long as the aspects of the mental that are revived are restricted to those that are susceptible to truth-evaluation ... then the behaviourists' criticism will be stayed (pp. 37-38).

These days the philosophers' understanding of cognition has now largely disappeared, even within philosophy of mind, where we witness many theorists grappling with issues that were initially considered to be non-cognitive.

Today many psychologists use the term *cognitive* in, what Shanon (1992) calls, a *phenomenological sense*, to simply refer to all those mental abilities and processes that make up the subject matter of cognitive psychology. Thus, cognitive psychology texts provide a reasonable window on this sense of *cognition*. The structure of modern texts has changed very little since the publication of Neisser's (1967) *Cognitive Psychology* beyond a steady trend in placing more attention on higher-processes such as problem-solving and reasoning[3]. Anderson (2000) provides a typical list of things cognitive in his widely used text *Cognitive Psychology and its Implications*: perception, attention and performance, knowledge representations, encoding and storage of memory, retention and retrieval in memory, problem solving, expertise, reasoning and decision making, and language. So this use of *cognition* serves as a kind of disciplinary shorthand, but unfortunately contains no explicit understanding of what it is for something to be cognitive.

Typically, when quizzed about the actual meaning of the term *cognitive* psychologists advert to describing the information-processing approach to understanding cognition (Shanon [1992] calls this the *theoretical sense* of cognition)[4]. On this reading cognitive phenomena are phenomena that are explained in terms of symbols (and, in connectionism, subsymbols) and their processing. That is, something is cognitive if it involves mental representations and their computation, however those concepts are understood. It is interesting to note that even modern psychological dictionaries shy away from explicitly saying what cognition is and encourage a 'theoretical sense characterisation'. For instance, the *cognition* entry in *The encyclopedic dictionary of psychology* (Harré & Lamb, 1983) tells the reader to "see language and cognition." (p. 90). Although the entry under *language and cognition* is mostly concerned with the relation between thought and language, it does mention that the modern use of cognition "is wider in scope than the study of thinking: a reasonable definition would refer more generally to the mental processing of information

---

[3] Sternberg (1998) provides a single exception to this rule. The structure of *The nature of cognition* is divided up into major conceptual themes, such as domain-generality versus domain-specificity, rather than the usual substantive topics.

[4] Some may argue that all theories of cognition, traditional and non-traditional, are information-processing theories broadly understood. I think that this really stretches the sense of information-processing well beyond that typically characterised within cognitive psychology. For instance, I would not consider ecological approaches (McCabe & Balzano, 1986) and dynamical approaches to cognition (e.g., Thelen & Smith, 1994) to be information-processing accounts. Chapter 2 of this thesis involves a critical analysis of, what I take to be, the mainstream information-processing model of cognition.

..." (p. 332). Bizarrely, *The Blackwell dictionary of cognitive psychology* (Eysenck, 1990) has no entry for cognition at all! It does however characterise *cognitive psychology* as "concerned with information-processing, and includes a variety of processes such as attention, perception, learning, and memory; it is also concerned with the structures and representations involved in cognition." (p. 61).

Another popular rendering of cognition is an *intermediate sense* (Shanon, 1992) where cognitive phenomena are understood as phenomena described at the explanatory level between the behavioural and biological levels. This level is usually thought of as an abstract, functional description of the operation of the brain. Such a view is most strongly associated with functionalist theories of mind (see Bechtel, 1988 and chapter 2 of this thesis).

None of these senses of cognition seem to get at the idea that cognition is to do with *knowing*. An etymological diversion may help us to sharpen our focus here. According to the *Oxford English Dictionary* the word *cognition* did not appear in the English lexicon until 1447, and it was not until 1651 that it was used in the philosophical sense as the action or faculty of knowing distinct from emotion and volition (Simpson & Weiner, 1989, Vol. 3, pp. 445-446). Prior to this *cognition* seems to have been a synonym for *understanding* or *insight* as the following selections from early texts indicate:

Illumynyd she is wyth clere *cognycyoun* In hyr soule (1447, Bokenham, *Seyntys*).

With conscience and perfit *cognition* of innocencie (1604, Wright, *Passions*).

I will not be my selfe, nor haue *cognition* Of what I feele (1606, Shakespeare, *Troilus & Cressida*).

(Source, Simpson & Weiner, 1989, Vol. 3, pp. 445-446).

This sense of *cognition* is etymologically related to the older English (Germanic) words *can* (to be able to), *ken*, and *know* and has much in common with the pre-Cartesian use of the word *mind*. Thus *cognition's* roots lie in the older 'commonsense' notions of *know* (I know how to milk cows), *can* (I can build a house), and *ken* (Writing a book is beyond my ken). Like the word *mind*, the early usage of *cognition* was connected with talk about a person's abilities and capacities rather than a type of reflective, internal mental activity.

I want to suggest that we should understand cognition in the old-fashioned sense as being about the abilities of humans and other animals to act. This may seem an odd move to make, because it ties cognition very strongly to the notion of behaviour, or at least the capacity to behave. Indeed, I will use the term *cognitive activity* interchangeably with that of *behaviour* where I understand behaviour in the non-behaviourist sense of "the self-initiated, purposeful movement of animals and humans." (Keijzer, 1997, pp. 11-12). Behaviour, in this interactionist sense, is not about mere movements, surface phenomena, a response to a stimulus, or an operant. Many modern psychologists are reluctant to use the term *behaviour* for fear of being branded as behaviourists, but as Hendriks-Jansen (1996)

notes, only a better theory of behaviour will put the ghost of behaviourism to rest. Behaviour, especially adaptive behaviour, is a complex phenomenon that encapsulates issues of knowledge, feeling, motivation, goal-orientedness, and, importantly, doing. I believe these concerns lie at the heart of what Shanon (1992) calls the *genuinely psychological sense* of the term *cognitive*. In this sense *cognition* refers to a phenomenon "only inasmuch as it pertains to meaningful behavior." (Shanon, 1992, p. 245). Keijzer's interactionist take on behaviour provides us with the kernel of an idea for appreciating cognition in this sense.

> Behavior does not only consist of movements, these movements are made 'in order to' bring about a specific consequence. On a longer time scale, these consequences are fed back to the behaving system (for example by selection in evolution or by reinforcement in learning). *The net, long term result is the existence of systems that maintain themselves over time.* (Keijzer, 1997, p. 12, emphasis added)

Keijzer's final sentence is an important one because it links behaviour with the activities that living things must engage in to stay alive. In chapter 6 I will examine a couple of frameworks that bring these concerns to the fore. For now it is enough to appreciate that I take cognition to be strongly bound up with an organism's capacity to coordinate its activities with the environment such that it can stay alive. One can see that meaningfulness is deeply grounded in such a relation because things (objects, events, actions) in the world will relate, to some greater or lesser degree, to the task of staying alive. Thus the ability to perceive in an adaptive manner is an ability to appreciate meaning as a kind of 'value perspective'. Therefore cognition, as it is understood here, relates primarily to the capacities living things possess for adaptively negotiating the world. The theoretical challenge of explaining how such behaviour is possible constitutes the central problem for cognitive science.

The Problem of Cognitive Order

Psychologists and their forbears have long been puzzled by the behaviour of people and other animals. This puzzlement comes from the fact that the behaviour of animals is qualitatively different from the behaviour of non-living systems. The Classical Greeks were aware of these differences and proposed that the *psyche* was responsible for them (Everson, 1991). The theories of Early Modern science developed by Newton and his contemporaries have had astounding success explaining the behaviour of the non-biotic world. However, their failure to do the same for the behaviour of animals, especially the higher animals like ourselves, has led a number of scientists in disciplines as diverse as physics (Schrödinger, 1945), biology (Ho, 1993), and psychology (Kugler & Turvey, 1987; Swenson & Turvey, 1991) to claim that the puzzles of human behaviour will require a more complex understanding of nature than that provided by orthodox Classical physics. Of all psychologists, Gibson (1979/1986, p. 135) was perhaps the most intensely aware of this problem:

The richest and most elaborate affordances of the environment are provided by other animals and, for us, other people. These are, of course, detached objects with topologically closed surfaces, but they change the shape of their surfaces while yet retaining the same fundamental shape. They move from place to place, changing the postures of their bodies, ingesting and emitting certain substances, and doing all this spontaneously, initiating their own movements, which is to say that their movements are *animate.* These bodies are subject to the laws of mechanics and yet *not* subject to the laws of mechanics, for they are not *governed* by these laws. They are so different from ordinary objects that infants learn almost immediately to distinguish them from plants and nonliving things.

Living systems like people exhibit a startling and complex form of ordered and meaningful (that is to say, adaptive) behaviour that is very rarely seen amongst non-living macroscopic objects and events. The problem of explaining how it is that living agents do this is what I call the *problem of cognitive order.*

The problem of cognitive order is an adaptation of social theorist Anthony Giddens' (1984) notion of the *problem of order*[5] – "the problem of how it comes about that social systems 'bind' time and space, incorporating presence and absence." (p. 181). Held and Thompson (1989b) provide a nice gloss on these complex sounding ideas.

Rather than thinking of time and space as abstract categories or as frameworks within which action takes place, they can be more illuminatingly thought of, Giddens suggests, in terms of 'presence' and 'absence' - terms which he borrows from Heidegger. Every interaction involves different forms of presence and absence. A face-to-face interaction typically takes place in a definite setting and endures for a definite period; the other person is 'present' both spatially and temporally. But social systems can become 'extended' in space and time, in such a way that the other is no longer immediately present. This time-space distancing (or 'distanciation', as Giddens generally calls it) has been facilitated by the development of new forms of transport and communication. The significance of the invention of the telegraph, to take one example, was that it separated the process of communication from the physical transportation of messages, thereby enabling individuals to communicate quickly at a distance, without sharing a common physical locale (pp. 7-8).

Although Giddens' focus is social theory we can understand 'problems of order' more generally as having to do with explaining puzzling phenomena in the light of an explanatory framework that does not 'allow' the existence of the phenomena in question. These puzzling phenomena often incorporate an action-at-a-distance principle where the system under examination is both affected by, and able to affect, things distant in time and space. In the absence of knowledge of the structures or mechanisms that make these phenomena possible, researchers are faced with a problem of how the system 'binds' space and time. This problem occurs because it seems that there exists no obvious mechanism for effecting the behaviour observed. Put another way, there appears to be an absence of a causal mechanism. For instance, to someone not familiar with a magnet, its power to influence a metal object without touching it may appear miraculous. The magnet's capacity

[5] Giddens has in turn borrowed this idea from Talcott Parsons.

to act at a distance contravenes the commonsense notion that a tool cannot be responsible for moving another object unless the tool is touching the object[6].

When applied to people and other living things the problem of order relates to the ways in which these agents can know, influence, and be influenced by things in the world distant in time and space. In other words, many animals do not seem to be stuck in the immediate here and now. They possess abilities to 'take account of' things that they are not currently in physical contact with. These abilities are central to the phenomena studied by cognitive psychologists and they seem to have one important feature in common: *they are all about adaptively modulating behaviour through being sensitive to the dynamics of the environment.* Reed (1996) nicely summarises this view when he writes "[f]rom an ecological point of view, in which knowing is not separated from living, cognition might best be defined as *an animal's capacity to keep in touch with its surroundings.*" (p. 169, emphasis added). Some phenomena, such as attention and perception, are essentially concerned with how people and other animals manage to know about, and thereby coordinate their behaviour with respect to objects, events, and properties in the *local environment*; that is, things with which the agent is in causal, but often not direct, physical contact. Other cognitive phenomena concern the ways in which animals modulate their behaviour with respect to things that exist in, what I will call, *distant environments* – environments with which the agent is not in direct causal contact. For instance, memory research involves understanding how past events and experiences shape current functioning. Problem-solving and reasoning often involve being influenced by anticipated, future circumstances. Language use also often involves mediating behaviour by using information about distant environments or by providing a way of redescribing what exists in the local environment. Unlike non-living things, people and other animals possess the power to sense the world about them, learn from previous experiences, make use of these 'memories', and, perhaps most importantly, anticipate the future. In a sense living systems are characterised by the fact that, to greater or lesser degrees, they 'stretch' themselves over spans of space and time incorporating and integrating the influences of distant things into their present activities. Unlike the particles of Newtonian physics, there is more to the behaviour of living things than the effects of collisions in the here and now[7]. These ideas

---

[6] The ancient Greek philosopher Thales assumed that magnets possess a *psyche* for this very reason (Everson, 1991).

[7] Things are, of course, not this simple. Some non-living systems, especially non-linear, far-from-equilibrium systems, exhibit very complex 'action-at-a-distance' behaviours that appear to be rather life-like. (The above magnet example is a simple non-linear, equilibrium system). Throughout history complex non-living objects and events have been understood as animate and agentive, possessed by spirits, gods, or other

seem to provide a promising avenue for making explicit the unifying concepts that underpin research into cognition. Our primary interest in basic cognitive research is in understanding how and why the behaviour of animals, and possibly other organisms, is so complexly modulated by environmental states and events. One useful way of characterising the problem of cognitive order is to produce an abstract map of the kinds of interrelationships that seem to exist between environment and agent.

Cognition as Time-Space Distancing

In this section I sketch out the variety of phenomena that exercise the minds of cognitive psychologists and attempt to show how they are all varieties of time-space distancing. At this stage I make no attempt to explain these abilities, although I occasionally use the ideas of various cognition researchers, cognitivist and non-cognitivist, to illustrate my claims.

Figure 1.1 provides a rough diagrammatic representation of the ways a literate, adult human (the typical participant in cognitive psychology experiments) 'cognitively stretches' over spans of space and time. To a first approximation it may help to think of the diagram as a map of the knowledge an agent has of objects and events that can influence its behaviour. The agent in question sits in the middle of this *time-space distancing map* (or TSD map): here the agent is temporally in the present and spatially in 'the here'. For the moment I will ignore the significant problems associated with the question of the duration of the present and rely on the intuitive notion that it consists of recent and soon-to-happen events rather than an infinitesimal point[8].

There are two primary areas of influence for the agent: the first consists of the objects and events that are 'copresent' with the agent; the second consists of the objects and events that are absent but often potentially knowable and thus influential. The size of the 'here and now' is demarcated by the perceptual abilities of the agent and its/their current effectiveness. Anything 'beyond perception' exists in the realm of the 'absent' and must be known by other means.

---

supernatural, but life-like, beings. Conversely, some living things, most plants for instance, seem more like non-living systems. In chapter 6 I address some of these complexities in greater detail.

[8] Both Gibson (1966) and Giddens (1984) have made the point that the present is not experienced as a point-like 'now'. Giddens (1984) notes that "... one might suppose that memory refers simply to the past-to-past experiences, traces of which somehow remain in the organism. Action then occurs in the spatiality of the present, drawing upon memories of the past whenever such are needed or desired. A moment's reflection will demonstrate the inadequacy of such a view. 'Present' cannot be said or written without its fading into the past. If time is not a succession of 'presents' but 'presencing' in the sense attributed to this by Heidegger, then memory is an aspect of presencing." (p. 45). The Heideggerian notion of presencing is similar to William James' notion of primary memory.

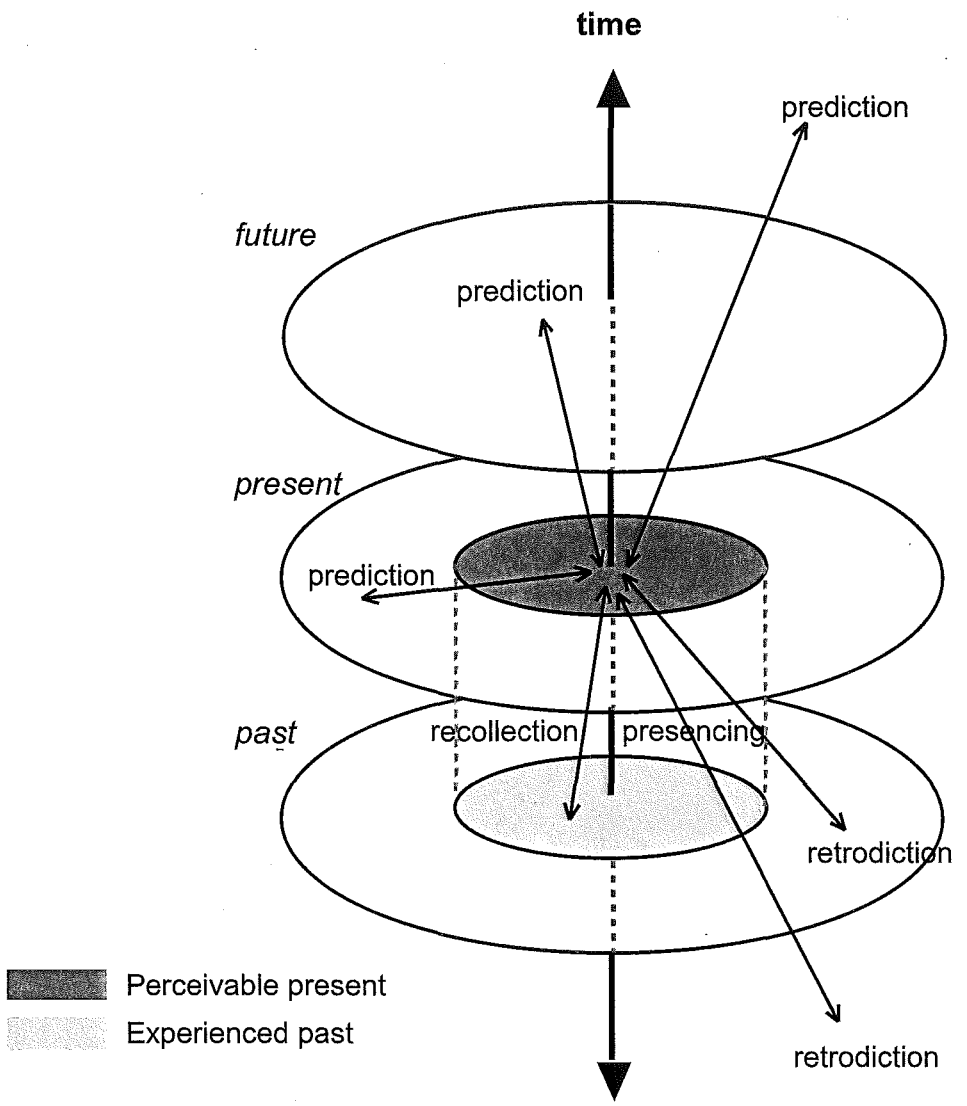

Figure 1.1 An Example of a Time-Space Distancing Map

Both the 'domain of the co-present' and the 'domain of the absent' feature in the problem of cognitive order. The former relates to what Clark and Toribio (1994) call the *compression* and *dilation* of the "unruly ambient manifestations" of objects and events - a distinction alluded to in the earlier work of Fodor and Pylyshyn (1988). The latter relates to Clark and Toribio's description of action in the absence of informative environmental signals. I will discuss these two ideas in turn.

Co-presence: Problems of Compression and Dilation

An agent is connected to its surroundings by its perceptual systems. Perceptual systems are used to 'pick up' information from the surroundings. Both *range* (how far they reach) and *acuity* (how finely they distinguish things) are important system factors relating to the utility (quality, usefulness) of the information that can be picked up. An agent possesses a

great advantage in being able to detect helpful or dangerous environmental states before it contacts them. Distal perception effectively enables the animal to 'see into the future' by being able to act early to avoid dangers or capitalise on opportunities (Christensen & Hooker, in press; Dennett, 1996; Smithers, 1995).

Thus, the *basic cognitive order problem* faced by an agent is how information can be detected without direct physical contact between the object or event in question and the perceiver's body. Turvey, Shaw, Reed, and Mace (1981) suggest that, at least with regard to *visual* perception, this basic order problem is central to the theories of J. J. Gibson:

> The intentionality of visual perception can work only by explaining how organisms can "come into psychological contact" with objects which they are not in physical, or more aptly, mechanical contact. Solving this problem of perceptual "action at a distance" is the function of Gibson's theory of ecological information for perception. (p. 242)

Typically this problem is understood by suggesting that the sensory systems of an agent make contact with the local medium as it is perturbed by the distal object or event (see Gibson, 1979/1986). Hearing involves a sensitivity to sound waves travelling through the air (and sometimes other media), vision consists of intercepting reflected light (photons), olfaction is the uptake of various volatile chemicals and so on. The action-at-a-distance powers of perceptual systems are thus explained by reference to the mediating effects of proximal causal consequences of distal objects and events.

But finding a pathway for the transmission of information from a distant object or event cannot explain how an agent responds *selectively* to types of proximal stimulation (see Keijzer, 1997, chap. 2). The major puzzle here is how to explain the ways an agent compresses and dilates 'input'. This problem encompasses two related sorts of phenomena. The first relates to the perceptual notion of *object constancy* (Humphreys & Riddoch, 1984), while the second involves an attempt to explain *conceptual equivalence.*

Object Constancy and Inconstancy

The problem of object constancy comes from the observation that a particular object or event presents the perceiver with variable proximal or peripheral stimulation (e.g., stimulation of the array of retinal cells in the eyes) when the object is viewed from different positions or in different situations. Simply put, this fact raises the question of how agents know that they are looking at (or act in the same way toward) the same thing when they see it at different angles, in different configurations (e.g., when partly obscured), and in different conditions (e.g., in a variety of lighting conditions). Object constancy is a compression problem simply because it involves understanding how an agent 'works out' how different types of stimulation arise from, and thus relate to, a single object (see Brunswick, 1952; Keijzer, 1997).

The computational vision literature does not explicitly address the possibility of an opposite sort of problem (what we might call a *problem of object diversity*), but Clark and Torbio's (1994) work suggests it is a possible phenomenon of interest. They note that compression problems (where agents treat different stimuli as equivalent) are mirrored by *dilation* problems (where agents treat similar stimuli as different). This sort of problem is not as easy to formulate as that of object constancy, and I suspect this has to do with the fact that it is a somewhat incoherent one. Presumably the *actual* difference between the two objects of identical (proximal) appearance has important consequences for the perceiver. For example, perhaps we dilate proximal input when we can tell a real apple from a wax replica[9].

Most cognitive scientists, including psychologists, adopt a sort of Kantian attitude to solving these problems, arguing that the perceiver's mind/brain contains mechanisms for ordering (squashing together for compression, prizing apart for dilation) the input stimulation in such a way as to either construct an internal canonical representation (e.g., Marr & Nishihara, 1978; Marr, 1982) or a manipulable, viewer-centred exemplar of the objects (e.g., Tarr & Pinker, 1990). These representations then provide the 'internal stimulus' for consequent action.

Ecological psychologists on the other hand claim that every object or event structures the appropriate informational array (the light for vision, sound waves for audition, etc.) in an invariant way no matter how or under what conditions it is observed thereby providing an unmistakable 'proximal stimulus' for object and event recognition (e.g., Gibson, 1979/1986; McCabe & Balzano, 1986). When an agent fails to perceive an object or event accurately this is not viewed as a consequence of faulty internal ordering mechanisms or the inappropriate use of stored knowledge, but rather as a consequence of the agent not being able to pick up sufficient information. Objects and events are dilated because they really do appear to be different, not because internal mechanisms separate them 'in the mind'.

Conceptual Equivalence and Nonequivalence

Clark and Toribio (1994) make a significant concession to 'direct perception' at this basic level, one that earlier researchers and theorists such as Ullman (1980) and Fodor and Pylyshyn (1981) were unwilling to make. These earlier researchers took the topic of perception to be the fixation of perceptual beliefs (beliefs about the things that people can

---

[9] Although surely if we could not notice a difference between a real apple and a fake one no amount of 'mental processing' would be able to reliably lead to an accurate distinction. I suspect the sort of dilation that must occur in such a situation is a 'higher-level', conceptual type of inference of the sort I discuss in chapter 8.

see, hear, etc., such as the belief that 'that woman standing quietly over there is a political reactionary'). Even basic perception of, say 'a person standing still' or 'a cliff with a long drop' were regarded as types of perceptual belief – situations where there is not enough information reaching the perceiver to specify these appreciations of the surroundings.

Research in situated robotics[10] (e.g., Brooks, 1991a, 1991b) has convinced a number of recent theorists, such as Clark and Toribio (1994), that it is possible that there is enough information in the world for agents to make basic perceptual compressions and dilations without requiring recourse to a mental representational explanation. However, they go on to argue that there exist *higher-level* compression and dilation problems ('real' perceptual belief problems) where the ambient physical manifestations of objects and events are far too different (compression) or similar (dilation) to provide the appropriate information for making functional distinctions[11]. They give two 'compression examples': the set of 'all valuable things', and the set of 'things that belong to the Pope'. Valuable things would consist of objects and events with such unruly ambient physical manifestations as BMWs, $100 bills, and friendships, whereas the Pope's belongings might include a Bible, a cassock and responsibility for the Roman Catholic church. Within cognitive psychology some concept researchers suggest that people often rely upon theories to understand class membership rather than mere perceptual similarity (Medin, 1989; Murphy & Medin, 1985). They argue that only the use of a theoretically-based strategy can explain the fact that children, money, photo albums, and pets all belong to a single category (i.e., 'things to take out of one's house in case of fire') (Medin, 1989). Clark and Toribio (1994) do not give a high-level version of dilation, but a possible example may revolve around a difference of functional perspective (or as Gibsonian's would say, a difference in affordance) of a $10 note lying abandoned at the side of the road and a $10 note sitting in a friend's wallet. Obviously both $10 notes provide similar ambient physical manifestations but the behavioural consequences of picking them up and putting them in one's pocket differ, because of the different contexts in which they are embedded.

Absence: Lack of Environmental Signals

So much for a person being able to have an 'ordered view' of their perceivable surroundings. As has been pointed out by many cognitive scientists, a person is never

---

[10] I discuss situated robotics research in more detail in chapters 3 to 7.

[11] Fodor (1986) makes essentially the same point when he argues that humans and other cognitively complex animals can respond to the *nonnomic* properties of objects as well as nomic properties. Roughly speaking, a nonnomic property is something that is not given by the so-called projectable properties of the object. For instance, I can respond to the fact that this computer is *my* computer although there is nothing about its projectable properties that indicates such a thing.

purely at the mercy of their current situation. To give just a couple of examples, Gibson's (1979/1986) distinction between direct and indirect (or mediated) perception and Corballis' (1991) distinction between praxic and non-praxic skills both relate to the idea that people can often act based upon knowledge they get from objects and events distant in space and time; that is, from objects and events that are unperceivable or absent (see also Kirsh, 1991). For Gibson, indirect perception comes about through using pictures, photographs, and other types of representation that 'carry some of the meaning' of the world elsewhere (and typically 'elsewhen'). Corballis (1991) suggests that praxic skills involve actions that are not "*stimulus-bound* - that is, actions that occur directly in response to environmental events." (p. 206). For him, praxis is characterised as "the organisation of purposeful, sequential actions in which spatial constraints imposed by the environment are minimal." (Corballis, 1991, p. 213). The ideas of both Gibson and Corballis fit nicely with Clark and Toribio's (1994) notion of types of behaviour that require information for which there exists no (direct) environmental signal or constant causal contact between the agent and the 'stimulus for action' (see also Chemero, 1998b). A simple example of such a situation is the change in behaviour evidenced by a person who has heard that a certain street corner is a haunt for muggers. There may thus exist no perceivable difference between a safe and dangerous street corner. Something else must be required in order to effect the behavioural change. What follows is a brief examination of the absent 'places' where this 'extra' knowledge may come from.

The Hidden Present

Real objects and events that are either hidden from the perceiver or out of range of their sensory systems can still be known in a number of ways. That is, there exist *stretching* or *time-space distancing mechanisms* that enable the knowledge of things in the 'hidden present'. One of the most obvious stretching mechanisms is *prediction*. Knowing what time of the day it is in a distant country prior to making a long-distance phone call is a simple example. The accuracy of this kind of stretching is of course limited by the predictability of the relevant events. Another way of obtaining information about things 'out of sight' is to be told about them by another person or warned by some sort or interpretable sign such as surveillance or communication equipment. Of these 'interpretable signs' Gibson (1979/1986) notes:

> Some optical instruments merely enhance the information that vision is ready to pick up; others ... require some inference; still others ... demand a complex chain of inferences. Some measuring instruments are closer to perception than others. (p. 260).

In these cases the reliability of the 'original direct perception' (e.g., the perception of the person telling you about the event, or the product of the resolving capacity of the instrument) of the event or object provides a bound upon the utility of this mediated

information. Hutchins' (1995a) study of U.S. Navy navigation crews provides a relevant example:

> The Officer of the Deck [OOD] tends to trust the navigation plot maintained on the bridge better than he trusts the plot generated in the Combat Information Centre [CIC]. This is because the OOD can come to the chart table, look at what the navigation team are doing, and talk to the quartermasters. In this relatively rich face-to-face interaction, an understanding can be negotiated. The work that went into the recommendation can be displayed and discussed. Such negotiation of meaning is difficult when one is dealing with CIC via the phone talkers. (p. 232).

Similarly, Vicente and Burns' (1996) investigation of how nuclear power plant operators monitor the state of their plants shows that the operators prefer direct knowledge of the state of components to 'second hand' knowledge from the reports of other workers, and the vast array of information displayed in the plant control rooms because "there are crucial differences between direct perception and indirect perception mediated by instruments. These differences arise from the fact that direct perception is based on information, whereas indirect information is based on external symbolic representations." (pp. 276-277). They note Gibson's (1979/1986) observation that for the bearers of 'second hand' information "the reality testing that accompanies the pickup of natural information is missing. ... The invariants have already been extracted. You have to trust the original perceiver." (p. 261). Trust (or lack thereof) thus constitutes an important constraint on 'indirect knowing'.

The Experienced Past

Past events and objects can also have an effect upon a person. Indeed knowledge gained from past experiences plays a vital role in all forms of time-space distancing. Things experienced by a person are often eligible for deliberate and conscious recollection and often involve the 'recovery' of facts and images. This is usually what people mean when they refer to memory and remembering. But the past also has a deeper effect in that experiences implicitly 'shape' or 'sculpt' the bodily skills of people regardless of whether a conscious recollection of the relevant learning events is possible. For instance, maintaining the ability to swim – a type of know-how or procedural knowledge – may stay with a person long after they have any explicit recollection of the places, events, and even learning procedures that were involved in gaining the skill. In a sense, memory is better thought of as the 'processes' where people's and other animals' conduct is regulated by their history of interactions with events and objects (Skinner, 1989; Maturana & Varela, 1988). Many cognitive psychologists have argued that there exist a number of different kinds of memory system and consequently that there are a number of different ways in which the past may modulate current action. One common distinction is that between *explicit memory* and *implicit memory* (Graf & Schachter, 1985; Schachter, 1987) where "[i]mplicit memory is revealed when performance on a task is facilitated in the absence of

conscious recollection; explicit memory is revealed when performance on a task requires conscious recollection of previous experiences." (Graf & Schachter, 1985, p. 501). Cohen and Squire (1980) have made a similar claim that learning can modulate our behaviour both *procedurally* and *declaratively*. Procedural learning is essentially motoric and skill-based whereas declarative learning involves the learning of 'facts' (and includes both *semantic memory*, memory for general knowledge, and *episodic memory*, memory for particular events [Tulving, 1972]). Although debate continues over whether either the implicit-explicit distinction or the declarative-procedural distinction better accounts for memory phenomena in normal and amnesic people, it is clear that there are two broad ways in which past events can modulate behaviour (see Glenberg, 1997). The first is a basic kind of skilful and unconscious shaping of perceptual and motor ability that is environmentally based and data driven. This notion of memory is roughly equivalent to what Heidegger called *presencing* (see Giddens, 1984). The second is an effortful and deliberate attempt to consciously 'retrieve' information and is person-initiated and conceptually-driven (see also Roediger & Blaxton, 1987). In the terminology of ecological psychology, implicit/procedural memory seems to be a sort of *direct memory*, and explicit/declarative memory seems to be a type of *indirect* or *mediated memory*. I will leave examination of these promising parallels for later. Suffice to note at this point that notions of memory are somehow related to the ways in which people and other living things can 'reach back' through time to events that mediate action in the present.

The Unexperienced Past

A person can also have knowledge of past events that they did not experience. This knowledge can be gained in two ways: through mediated memories that are memories of indirect perceptual activity (e.g., remembering that you had been *told* that the Treaty of Waitangi was signed in 1840), or through *retrodiction.* Retrodiction is simply prediction of things past but, unlike prediction, retrodiction is usually reliant on the existence of some sort of set of 'archives' to act as a basic contact with the past. Often these archives may have actually been present in some form at the time of the event being retrodicted (e.g., geological strata, fossils, archaeological findings, original texts and so on). These archives act as a *sign* or *representation* of things past and their utility and accuracy is both limited by the resolution and quantity of the information that they can carry and the ability of the person to 'pull out', or interpret, that information. Archives may also have been constructed after the relevant past event from the author's direct perception or upon their readings of other archives. In many forms of retrodictive activity there exist long chains of interpretation and 'reconstruction' of the past that incorporate numerous interpreters and archival sources. An undergraduate's textbook and lecture notes are prime examples of such processes. Through the use of authored archives a student of, say, astrophysics, can

reach far back in time and light years through space, in order to make authoritative statements about such untouchable phenomena as black body radiation and black holes.

The Future

Daniel Dennett (1996) reminds us that the "task of the mind is to produce future" (p. 57). And more specifically:

> The mind is fundamentally an anticipator, an expectation-generator. It mines the present for clues, which it refines with the help of materials it has saved from the past, turning them into anticipations of the future. It then acts, rationally, on the basis of those hard-won anticipations. (pp. 57-58).

The future is a domain that places interesting bounds upon action in the present. Our expectations, anticipations, and predictions of future events and objects are in large part responsible for our actions and activities in the present. Notably, there is no notion of 'direct prediction' - the ability to actually 'see' the future is reserved for those who claim they possess paranormal or time-travelling powers. Because all prediction is an indirect cognitive activity it is vulnerable to inaccuracy, unreliability, and confabulation. But not all prediction is equally untrustworthy. Prediction effects can range from the scientifically respectable (e.g., there will be a high tide at 6 a.m., so put the fishing gear in the car for an early start tomorrow) to the speculative and fantastic (e.g., my horoscope said to be careful about dealing with large amounts of cash today, so I'll put off buying the car until tomorrow).

Modern human life is typically a future-oriented activity. The consequences of actions are foremost in our thought and are 'built-in' to the discourses, institutions and organisations that we inhabit. Saving money in a bank, taking out a loan, having a job, and having a family all involve actions in the present which commit us to future actions, responsibilities, and consequences. Often we are chastised when we ignore or overlook the consequences of our actions by people as diverse as teachers, parents, governments, social commentators, and environmentalists. The future thus affects our present behaviour in interesting ways although, as Skinner (1989) correctly notes, the future only 'works on us' insofar as we *expect* things about it, and we only expect things because of a *past* history of particular types of interactions with our surroundings. However, this does not mean, as Skinner supposes, that "[w]e do what we do because of what *has* happened, not what *will* happen." (p. 14), unless, by such a remark, Skinner simply means that we do not directly perceive the future. The future is by no means fully determined, but our lives are fundamentally organised to exploit those parts of it that are likely to occur. Within cognitive psychology several research domains deal directly with the modulating behaviour with regard to the future. Judgment and decision-making research studies the ways in which people plan to act and problem-solving research examines the ways in which people move from a current state to a desired future state of affairs.

The Hypothetical

There exists one last 'space' from which a person can gain information for their current action. I call this the *domain of the hypothetical* (it is not represented in figure 1.1 above). By hypothetical objects and events I simply mean invented, imaginary, or (currently) fictional representations of states of affairs *that are acknowledged as such.* This distinguishes a 'hypothetical' from a prediction or retrodiction of a state of affairs. The latter are *intended* to capture information about things going on in the perceptually inaccessible world. I do not wish to draw too hard and fast a distinction here as many predictions are intended only as inaccurate possibilities that approximate truthfulness. This is, of course, an acknowledgment of the potential unreliability and inaccuracy of the procedures for making the predictions in the first place. In sum, the key point here is that our attitudes toward representations (i.e., whether they are real or unreal, possible or impossible) determines how they are likely to affect our current actions.

It may seem that acknowledging that a hypothetical representation has no real referent (whether in the present, past, or future) would be grounds for arguing that it should have no bearing on our actions. All adaptive activity should be based upon real or likely states of affairs, not things that we *know* are *not* the case. To act based upon a fictional reading of something would constitute a potentially dangerous strategy. But hypotheticals do have important and adaptive effects on the things that we do in the present. For instance, philosophers and scientists use hypotheticals when they engage in *counterfactual reasoning*[12] to work out what *might* have happened had some, say experimental, conditions been different. And what *might* have happened may provide us with knowledge about the world that is as useful as knowledge derived from what has actually happened. This sort of reasoning is not just central to the design and conduct of scientific research (if A really does cause B all by itself, an experiment that screens off A should not result in B), but also to a lot of everyday problem-solving where we plan to do things differently in future given the outcomes of particular experiences. For instance, on experiencing traffic problems taking one route to work you may reason that next time (a future action) you will take an, as yet, untravelled alternative route. This is a prediction of a future state of affairs, but it is a prediction based upon what *may* have happened had you done things differently earlier. It is not simple inductive reasoning from past experiences to future experiences, but a more complex version of reasoning that involves forays into hypothetical space where representations of experiences past are decomposed and then recomposed in accordance with a deeper understanding of underlying generative mechanisms. This sort of *existential*

---

[12] Crane (1995, p. 56) describes counterfactual reasoning when he notes that "... if we believe that A caused B, we commit ourselves to the truth of the counterfactual, 'If A had not occurred, B would not have occurred'."

abductive reasoning (see Haig, 2000) relies on what Annette Karmiloff-Smith (Karmiloff-Smith, 1992, 1994; Clark, 1993; Clark & Karmiloff-Smith, 1993) calls *representational redescription.* Broadly speaking, representational redescription refers to an ability to disembed implicit knowledge from its grounding in real-world, procedural activities and use it to facilitate manipulation and interaction with other knowledge. This enables a person to creatively combine different ideas, images, and concepts in a compositional manner. Much more will be said about the details of these mechanisms in future chapters. For now the important point is that the 'inhabitants' of hypothetical space feature centrally in the determination of actions in the present via the time-space distancing mechanism of representational redescription.

Conclusion

In this chapter I have set up a framework for the chapters to come. I believe that, just as the Ancients suspected, a genuine 'psyche-ology' is not so much a study of human mental processes but rather the *study of life*; of how it is that living things manage to use and negotiate the environment so that they stay alive. For too long the question 'what is cognition?' has been answered with the reprise 'see information processing'. Time-space distancing, on the other hand, gives us a relatively theory-neutral description of the issues underlying our use of cognition-related terms. The problem of cognitive order and the notion of time-space distancing will thus serve as a framework for investigating current cognitive science research in Chapter 2 and an alternative embodied and situated approach in the rest of this work.

2. Cognitivism and its Problems

We have all seen one social or human science after another - psychology, sociology, economics - come under the sway of some fad. In the United States such fads were more often than not the product of a reductionist idea of what it means to be "scientific." The idea that nothing counts as a contribution to "cognitive science" unless it is presented in terms of "mental representations" (and these are described "computationally") is just another case of this unfortunate tendency.

Hilary Putnam (1988, pp. 55-56).

Introduction

In this chapter I critically examine the primary assumptions that constitute the cognitivist approach to understanding cognitive activity. At the core of cognitivism is the basic belief that the mind is an in-the-head, physical, rational device (Bem & Keijzer, 1996). This starting point enforces the following constraints on the cognitivist enterprise: naturalism, rationality, and localised control. *Naturalism* refers to the belief that a materialist understanding of mind cannot make reference to semantic or intentional terms, for these are the very concepts that cognitive theorists attempt to explain (see, e.g., Crane, 1995)[13]. Thus naturalism demands an explanation of mental notions in physical or causal terms. *Rationality* is the belief that the mind is somehow endowed with the ability to execute bodily movements reliably and validly[14]. In other words, most of the time the mind accurately assesses an agent's surroundings and uses these assessments to co-ordinate the agent's body with the environment in an adaptive manner. *Localised control* derives from the assumption that the mind, being a device, has to be somewhere. For the materialist this location is nearly always taken to be the brain or, more broadly, the central nervous system. The mind thus forms a subsystem within the body and its function is to control its movements in the world in an adaptive manner.

Given these constraints, the basic task of cognitive science becomes the search for a likely explanation of how a brain can realise the rational device of the mind. Cognitivism provides a solution in terms of a framework of three interlocking assumptions: a commitment to an *information-processing* or *sense-model-plan-act (SMPA) architecture*; a commitment to a *representational-computational explanatory approach*; and the use of *formal task description* as the primary method for deducing the kinds of processes and operations that underpin cognitive activity.

---

[13] This is a more restricted sense of *naturalism* than that generally used in the philosophy of science.

[14] This is a broader sense of rationality than is used within philosophy and cognitive science. There rationality is used in the more restrictive sense of the use of reason or logic to solve a problem.

Cognitivism holds that people and other animals are *information-processors* in the sense that they make their way about the world by receiving stimulus information from their local environment and then transduce it into an internal informational code. The information in this internal code is manipulated and transformed in order to produce a useful model of the environment – a model that lets the agent know what is in the world, and where the agent is relative to those things. It is often argued that this model cannot be built entirely from stimulus information and thus must be supplemented with information from memory (considered to be a kind of database of extra supporting information, e.g., the basic concept of *schema* used in cognitive psychology). The agent knows where they are situated in the environment once the model has been constructed. The subsequent stage of processing then involves deciding what to do next. This is often thought to involve a process of internally simulating different courses of action (using the model as an internal off-line surrogate of the real-world) to see what the most effective actions will be. Ultimately a plan for action is constructed. Finally this plan needs to be translated into a 'motor program' – a series of instructions telling the body how to produce the desired behaviour. In sum, cognitivism is committed to, what a number of researchers call, the sense-model-plan-act (SMPA) framework (Brooks, 1991b; Smithers, 1995).

The SMPA idea is complemented by a theory of how these different 'stages of processing' are carried out in an agent's mind/brain. We can call this the *representational-computational theory of cognition*. In very general terms this is the claim that the brain is a kind of computer and that mental information is realised in the neural substrate in a manner similar to the way that information in a computer is realised by the computer's hardware. Thus, internal information (or knowledge) is thought to be representational (internal states of the brain represent various concepts, features, and properties). These representations are manipulated and transformed computationally; that is, according to certain kinds of rules. This view of the relation between the brain and mental activity may seem to be relatively straightforward, but fact numerous books and papers have been written trying to make sense of it. Philosophers have constructed many candidate *theories of mental representation* and *theories of content* (Von Eckardt, 1993). Theories of mental representation aim to describe how some physical objects (e.g., a neural ensemble or a state of a transistor) can represent other things. Theories of content are attempts at explaining how these representations get their particular representational content. Often these two projects are tightly interconnected in the literature that discusses the plausibility of the representational-computational theory of mind (These issues are discussed in more detail in the following section *Theories of representation*). Without going into the gory details at this stage, the upshot of this research is that nobody has constructed a widely endorsed, adequate theory of representation thus far (Grush, 1997). Indeed a number of philosophers

seem to be losing interest in the problem altogether. In sum, the notion of representation is currently vague at best.

The third key assumption of cognitivism is a methodological one aimed at providing a way of working out just what kinds of representations and computations are necessary for carrying out a particular cognitive task. The answers to this kind of question have always been at least partly empirically based. Many cognitive psychology experiments have attempted to use chronometric analyses to demonstrate how many information-processing operations are necessary to complete certain tasks (the usual reasoning is that if a task takes longer than another related task then it involves 'more processing'). However, underpinning this kind of research is a guiding belief in the usefulness of analysis using *formal task descriptions* (a.k.a. functional task analysis, decomposition by function) (Hendriks-Jansen, 1996). Basically speaking this is the strategy of hypothesising that cognizers, or cognitive subsystems, can be understood as having to accomplish particular tasks, and that the mechanisms underlying cognitive activity can be discovered using a logical or algorithmic breakdown of these tasks (the behavioural functions of particular activities) into a series of subtasks. Each subtask is thought to be implemented by a particular component of the agent's inner machinery[15].

In the rest of this chapter I will examine each of these assumptions, particularly the representational-computational theory, in more detail and show how they constitute a mutually-reinforcing set of arguments and ideas. That is, I want to suggest that in utilising the concepts of representation and computation there is an implied commitment to viewing cognizers as information-processors and a similar commitment to viewing the task of the cognitive theorist as the production of formal task descriptions that can be mapped on to the neurocognitive architecture[16]. I will also describe some of the difficulties associated with these different assumptions. In particular I focus on the idea that cognitivism's set of interlocking assumptions force on us a guiding methodological picture that is antithetical to an ESD interactionist view of cognition. Interactionism and representationalism cannot be easily reconciled without fundamentally reworking both the standard and intuitively attractive notion of representation and the useful elements of the interactionist approach.

---

[15] This component need not be a neurobiologically independent entity. Neural resources can be shared amongst different functional systems. That said, however, the equation of, at least some functional systems with specific neural systems is widely assumed.

[16] A number of authors have recently argued that the three cognitivist assumptions do not have to be understood as naturally implying each other. In particular it is argued that we may endorse many of the challenges of ESD interactionism, but preserve some kind of role for representation and computation (e.g., Clark, 1997; Markman & Dietrich, 1998). I will have more to say about these hybrid positions in coming chapters.

This kind of argument stands in contrast to many of the debates over mental representation in that it does not seek an ontological argument that aims to knock down a particular view of representation (although I do not presume that such a task is impossible). Instead I want to claim that, as cognitive research increasingly pays attention to issues of embodiment, situatedness, and distribution, the prominence of representational and computational issues will fade.

Using Representation and Computation to Explain Time-Space Distancing

At the heart of cognitivism is the idea that the agent is separated from their environment and that, in order to act adaptively in that environment, they must have access to knowledge about the environment. This assumption seems to demand a representational answer to the problem of cognitive order. The problem of cognitive order, it will be recalled, refers to the need to explain how an agent 'hooks up with' elements of the world beyond its skin in an adaptive manner. Cognitivism holds that this time-space distancing occurs because many animals are, or contain, mental representational systems. The notion of mental representation derives from an understanding of external and public representations such as pictures, signs, flags, texts, manual gestures, and spoken utterances, as well as natural signs such as clouds that signify impending rain, spots that signify measles, and smoke that signifies fire. Not surprisingly then, a number of modern cognitive theorists have used C. S. Peirce's (1931-1958) comprehensive treatment of the notion of representation to unpack the notion of *mental* representation (e.g., Deacon, 1996, 1997; Von Eckardt, 1993). Given this relatively widespread endorsement I will use Peirce's framework to provide us with an idea of the kinds of relations and structures that must hold for something to be a representation.

Peirce's View of Representation

Peirce distinguishes three different types of representation: icons, indices, and symbols. Icons are things that share certain qualities with the things they represent (their object). The small paint squares on paint charts are icons of paint (Peirce called these types of icons *images*). A road map is an icon of some terrain as there exists a mapping between aspects of the world and the map (a *diagram* in Peirce's terminology). Indices are things that have some sort of real causal or spatial-temporal connection with the thing they represent. A rap on the door is an index of the presence of someone at the door. A weathervane in a certain position is an index of wind direction. A pointed finger or an utterance such as 'now' or 'there' are also indices of the thing or things 'pointed at'. Symbols include all arbitrary human creations that are 'about' things and that have their meaning established by means of convention. They include words spoken and written, flags, road signs, gestures and so on (for a more recent, but similar, view see Noble & Davidson, 1996, p. 5).

Although, in some respects problematic, Peirce's distinctions are useful for picking out the sort of things that may be called external representations. In all three cases Peirce is clear that a representation has three integral relations. Any 'thing' that does not possess all three is not a representation. A representation is a representation when there exists a *bearer* (or sign), an *object* (or referent), and an *interpreter*. A *bearer* is simply the physical form that a representation takes. Crane (1995) distinguishes between the *medium* of the representation (the stuff that realises the representation such as sound waves, ink on paper, a magnetic disk, fabric or wood) and the *vehicle* of the representation (the form of the instantiation, such as natural language or a picture or flag symbol).

An *object,* or referent, is simply the thing or things that the representation is about. Many theorists argue that it is important to distinguish a representation's *object* and *content*[17] (e.g., Smythe, 1989). Objects are typically understood to be the 'things in the world' that are represented. The content refers to the way in which the object is construed in the representation. The object of a photograph of a tree is the actual tree (whether it still exists or not) whereas the content of the photo is the aspect of the tree captured by the configuration of the bearer. This distinction hints at the many complex problems associated with the notion of a representational object (does a picture of a unicorn have an object or merely content?).

An *interpreter* is someone for whom the representation has meaning. Peirce called the 'mental effect' of a representation an *interpretant.* Some writers have taken this claim to be equivalent to the idea that external representations are understood when they cause a thought or mental representation. It is clear that without someone interpreting a bearer nothing is represented (if no one ever saw a particular set of deer hoof prints in the mud in a forest then they will never have actually represented anything). Yet it is always *possible* that someone could have interpreted the hoof prints as a representation of the presence of some deer. Von Eckardt (1993) notes that "[v]irtually anything is a possible representation for Peirce, but only certain things are actual representations." (p. 153).

Since Peirce's analysis of representation deals exclusively with external representations it does not explicitly provide a foundation for discussing the nature of mental representations. More recently theorists such as Dretske (1988, 1995) have attempted to recast the ideas of previous philosophers such as Peirce in order to make sense of the notion of mental or internal representation. Roughly speaking, Dretske accepts Peirce's analysis of the major aspects of a representation (the bearer, object, and interpreter) and takes account of them in his taxonomy of representational entities. He distinguishes three kinds of representation: conventional public representations, natural signs, and, what he calls type III

---

[17] Other terms used in making this distinction include *sense* versus *reference* and *intension* versus *extension.*

representations. The third sort of representation, Dretske contends, describes physical entities that can do the work of mental representations. Dretske's *type I representations* are equivalent to Peirce's symbols in that they are representations that gain their representational power via convention. According to Dretske type I representations do not have intrinsic representational power. Many, perhaps most, cognitivist theorists take type I representations to be grounded in mental representations. They are viewed as external expressions of internal thoughts and concepts that have found a high degree of stability across cultural groups due to various processes of social learning and transmission. For instance, Davies (1995) argues that linguistic aboutness derives from the intrinsic intentionality of 'attitude aboutness'. Attitude aboutness refers to the aboutness of human mental states such as beliefs, desires, hopes, wishes and intentions (the so-called propositional attitudes). Davies suggests that the aboutness of our propositional attitudes is more semantically fine-grained than that available in public language.

Type II representations are what Dretske calls *natural signs* (after Grice, 1957, see Bem & Keijzer, 1996; Davies, 1995, refers to the 'intentionality' of natural signs as *indicator aboutness*). These roughly correspond to Peirce's indices in that they gain their representational power from objective rather than conventional relations with the world. Well known examples of natural signs include smoke indicating fire, annual rings indicating the age of a tree, and small red spots indicating measles. Hendriks-Jansen (1996) notes that the notion of natural representation is not really representational in a strict sense.

> It is certainly not the function of tree rings to indicate the age of a tree, nor, presumably, did they indicate such a thing until human beings learned to interpret them. Tree rings are the result of variations in growth rate over the seasons. They have no natural function and no natural "users" of their representational potential. (p. 40).

In an important sense type II representations are not *'natural'* representations at all, their capacity to be *used* as representations notwithstanding.

Dretske's *type III representations* are similar to natural signs but they *do* have natural users. That is, they are natural states of the world that are objectively related to other states of the world and they are used by the system of which they are a part because of this natural relation. These representations with natural users possess what Dretske refers to as *functional meaning* rather than the natural meaning of type II representations. Dretske introduces the term *functional meaning* to describe the idea that a type III representation serves a function or purpose within the system that it is embedded. If a component has a function, argues Dretske, then it can *mal*function, and if the function of a representation is

to represent then it can *mis*represent[18]. For instance, a petrol gauge (type III) represents the amount of petrol in the tank because that is its function (it has been designed to convey that information), whereas the annual rings on a tree do not exist in order to convey the age of tree to any natural user. Of course the petrol gauge example suffers from the fact that a gauge is a human artifact and is thus really a sort of type I representation because its interpretation and use requires appropriately enculturated humans. The car itself does not use the covariance to do anything. Dretske tries to provide a better example by appealing to the workings of a species of sea-going bacteria that possess internal magnetic structures called magnetosomes. The bacteria 'use' their magnetosomes to avoid toxic, oxygen-rich waters near the surface of the sea (i.e., they have an important biological function within the bacteria and thus have functional meaning) and Dretske wants to suggest that the magnetosomes thus represent oxygen-rich or toxic waters. We will examine some of the difficulties associated Dretske's analysis later on in the section *Representations as Indicators: Causal Theories of Content.* At this stage it I merely want to indicate the ways in which, at least some, cognitive scientists try to conceive of representations as natural intentional entities. Not all philosophers of mind consider Dretske's notion of type III representations to be sufficient for making sense of the broad range of intentional mental state terms used within psychology and the other cognitive sciences. Davies (1995), for instance, seems to suggest that Dretske's type III representations merely possess 'indicator aboutness' but that the aboutness of propositional attitudes (beliefs, desires, and so on), experiential aboutness (the intentionality of our perceptual experiences), and subdoxastic aboutness[19] (the aboutness of unconscious, subpersonal psychological states) possess different, possibly additional, qualities to those that make up type III representations.

Representations are thus widely thought to consist of bearers that carry informational content and that are used by an interpreting system. Such entities can enable a representation-using system to engage in a number of important activities.

What Representations Can Do

Mental representations are supposed to accomplish time-space distancing in the same way that external representations do. In this section I will examine the things that external

---

[18] Later on in this chapter I argue that a malfunctioning type III representation does not constitute a case of misrepresentation sufficient for cognitivist needs.

[19] Davies (1995) argues that subdoxastic aboutness differs from indicator aboutness because the former "allows for the possibility of misrepresentation." (p. 372). It is possible, then, that Davies may accept that Dretske's type III representations, which Dretske considers capable of misrepresentation (see the following section about *Representations as Indicators*), *could be* candidate entities for enabling subdoxastic aboutness. This is important because Davies' subdoxastic states seem to be equivalent to the subpersonal mental representations central to most modern cognitive scientific theory.

(public) representations make possible: time space distancing, making the abstract concrete, and making mistakes possible.

Representations can make the distal proximal. Things at a distance in time and/or space can have effects on people (interpreters) who are not co-present with them via the mediation of representations. High-tech satellite transmissions and the internet, for instance, can produce real-time representations of states-of-affairs occurring thousands of kilometres away. Books and letters do a similar sort of thing but delay contact. The fact that the bearers of representations can remain interpretable for great lengths of time introduces the possibility of contact with the past although, of course, this relies on the capacity of the interpreter to interpret the bearer (think of archaeologists faced with an artifact inscribed with an unknown ancient symbol system). And bearers are not restricted to things created by people or other animals. Geologists for instance can 'read the Earth' to recover something of the events of the past before humans, and indeed before life, existed on the planet.

Representations also provide us with the tools for making concrete imaginary or hypothetical states of affairs. We can draw a picture of a unicorn or tell a story about pixies that can affect an audience (even when that audience knows that the things represented are not real). Representations can also stand for non-imaginary things for which there is no concrete referent such as the 'number 3' or the 'species of cats' or 'justice'.

This brings us to a third important effect of representations. If we can represent things that are hypothetical or imaginary (even in the cases of prediction or retrodiction when we hope that our representational constructions will be, or were, real) then there exists the possibility that objects may be *misrepresented.* When an interpreter takes a representation as referring to something that does not exist (e.g., telling someone that there is a burglar in their house when there isn't) there exists the possibility that the interpreter will act upon the misinterpretation producing maladaptive (dangerous, embarrassing or incomprehensible) behaviour.

How Representations Do It

Representations make possible time-spacing distancing, 'concretisation' and mistake-making because they possess the properties of *localness*, *surrogacy*, and *persistence.* Representations are *local* in the simple sense that they have their effects on people by being in the same place as the reader or user of the representation. A 'Beware of the Dog' sign provides me with information about Sam even when Sam is hidden quietly behind the fence (i.e., unperceivable). Thus, we do not need to appeal to 'occult' or action-at-distance causation (e.g., my detecting Sam's presence by ESP) to understand how a person can have information about things in the world that are distant from the perceiver – that is, not obviously causally connected to the agent in some sense (see the section below on versions of representation hungriness). Newell (1990) makes this clear when he writes that:

> It is a law of nature that processing in the physical world is always local, that is, always takes place in a limited region of physical space. This is equivalent to there being no action at a distance, or somewhat more specifically, that causal effects propagate with an upper velocity of *c*, the speed of light in vacuo. Consequently, any computational system ultimately does its work by localized processing within localized regions in space. What guides or determines this processing task must then also be local. (p. 74)

Newell considers symbols, or more accurately symbol *tokens*, to be things that cause us to access and retrieve distant knowledge rather than being informative in their own right. Newell's notion of a symbol is somewhat like that of a variable in a computer program. A variable is assigned a value at some point in the operation of the program. When the program executes a procedure that involves the variable the variable's current value is accessed and retrieved from elsewhere in the computer's memory.

Representations thus can serve as *surrogates* or *proxies* for environmental information. I can be scared of Sam and behave appropriately without ever having laid eyes on him. Recent views of representation in cognitive science have focused on the idea that we can tell that something is a representation when an entity 'stands in for' aspects of the environment that are not always reliably present (Bechtel, 1998; Clark & Toribio, 1994; Grush, 1997; Haugeland, 1991).

Finally, many representations possess the property of *persistence*. They stay the same, or, more generally, they remain potentially informative, over time. Their persistence is largely determined by the physical nature of the representation bearer. Representational persistence is important because it provides an avenue for storing information of the past. In other words, persistence constitutes a sense of memory. This allows the user of the representation to let the representation 'carry the memory-load' thereby freeing the user to devote their cognitive resources to other activities. If a representation is transportable (e.g., a list of appointments in a diary) then it can be used as a mobile information carrier and it can be consulted in a whole variety of environments, as long as the environments support the use of the representation bearer. Thus, mobile representations permit the use of information in multiple contexts, and this capacity permits a certain level of decontextualisation that would not be possible without them. Consider the following situation: a friend tells me a list of important phone numbers which I could not possibly remember unaided. If I go home without recording them, and my friend is elsewhere, I can not make use of the phone number information; the 'environmental stimulus' is not available as a resource for action. If I write the numbers down in my diary and take it home with me, a surrogate of the environmental stimulus comes with me. The diary entry permits me to use the information in the absence of the environmental stimulus provided by my friend; I can perform actions 'out of context'. Moreover, I could not do this unless I had the representation and the ability to use it.

How Representations Solve the Problem of Cognitive Order

If there exists an internal and mental equivalent of external representations, cognitive scientists can explain a huge variety of human and animal activities that would otherwise appear puzzling or at least recalcitrant to explanation under the environmental determinism assumed in behaviourism. If we can make sense of the notion of a mental representation we can avoid the 'right here, right now' implication of the behaviourist's situational determinism. In other words, mental representations, like their source external representations, can account for the ability to behave adaptively when faced with *representation hungry problems* (Clark & Toribio, 1994). These are situations where there is no reliable environmental signal of important behavioural determinants (see Chemero, 1998b for a critical discussion of this idea). The mental representations stand in for the perceptually absent things by serving as an internal surrogate for what is missing externally. Mental representations can thus account for behaviour that is based upon belief rather than what is perceived, and thus (mis-) behaviour that occurs because of the misrepresentation of a situation. And because mental representations are viewed as neural states in animals, and similar underlying internal, physical states in other potentially intelligent agents (robots, Martians), they are always close at hand. In a sense the notion of mental representations as being inside a cognitive agent explains behaviour in the absence of external stimuli by suggesting that the 'stimulating' world is to a large degree carried about inside them.

Versions of Representation Hungriness

In chapter 1 I described several subproblems that together constitute the problem of cognitive order. These subproblems include what I called the basic problem, object constancy, conceptual equivalence, the hidden present, the experienced past (temporal stability), the unexperienced past, the future, and the hypothetical. There exist important differences of opinion within modern cognitive science over the degree to which solutions to the various subproblems mentioned previously require appeal to the notion of mental representation. These differences of opinion largely revolve around the sort of thing the various theorists take a mental representation to be.

Orthodox cognitivists believe that all of the subproblems are solved by mental representations. I will call this view the *close range version of representation hungriness* because its advocates argue that even things that are only slightly distant from the mind/brain have their cognitive effects on the agent via some form of mental representation, at least in higher animals like ourselves. Representations in this sense are any sort of internal (neural) state configured by informational stimuli (either environmental or internal knowledge) that are used to mediate behaviour. Fodor and Pylyshyn (1988) and Lloyd (1989) might be close range advocates.

An increasing number of researchers believe that we do not need to appeal to the notion of mental representation in order to explain an agent's (often complex) behaviour in response to information available in the local (i.e., perceivable) environment. Instead they argue that an internal state is only truly representational if it provides, or could provide, information in the absence of environmental stimuli[20]. Thus, a 'neural mapping' of something in the *local* environment would not be considered to be a representation but perhaps only a *presentation* (Grush, 1997). Representations are thus a form of internal model of what *may* or *could* be. This view I call the *medium range version of representation hungriness*. Clark and Toribio (1994) and Grush (1997) fit nicely into this category.

Advocates of the *long range version of representation hungriness* suggest that many cognitively complex activities are possible without recourse to the notion of representation and that only humans who have acquired representation-using skills can aspire to truly long-range forms of time-space distancing. These kinds of representation are not internal or mental representations in the traditional cognitivist sense, but rather are private versions of 'external representations'. Gibson (1979/1986), Noble and Davidson (1989, 1996), and possibly Vygotsky (1930-1935/1978, 1934/1986), can be seen as endorsing this sort of version of representation hungriness in different ways. Bechtel (1993a) and Hutchins (1995a) may be other advocates. It is also the view that I will develop in what follows (especially in chapter 8). The long range view is not what I would consider a cognitivist view of cognition because it rejects the notion of mental representations as 'things in the head'. This is not to say that there are not interesting facts of biology (especially neurobiology) associated with this view.

Aspects of the Problem of Cognitive Order

In the earlier discussion of the subproblems of cognitive order (in chapter 1) I hinted at the ways cognitivist researchers use the notion of mental representation to provide solutions to these problems. I will now take time to briefly tie these ideas into the current examination of the notion of mental representation.

According to some the basic problem of how the distant world affects an organism is solved by the fact that the environment is represented by the activation of arrays of cells on an organism's sensory surfaces. In Marr's (1982) theory of vision the basic resource for

---

[20] Clark (1997; Clark & Grush, 1997) argues for a weakened form of the 'stand in' version of representation (see 'Representations as simulations' below). In the 'stand in' view an entity is only a representation if it can and does stand in for an absent object. In Clark's (1997) view there exists a weaker kind of representation that cannot be decoupled and used off-line but which has the evolutionary function of standing in for its object. This makes Clark, at least partly, a close-range advocate. See Chemero (1998b) for a critique of all 'standing in' theories of representation.

seeing the environment is the continually changing pattern of stimulated retinal cells in both eyes (what is often called the *peripheral stimulus* and what Marr refers to as an *image*). Many researchers, especially those in the neurosciences, are happy to refer to these patterns as representations.

The problem of *object constancy* is solved by the computational generation of a mental representation common to a variety of activated stimulus arrays. That is, an object is known as such because there is a many-to-one mapping between peripheral stimulation and an exemplar or canonical representation.

Conceptual equivalence is handled in a similar manner to object constancy, the difference being that conceptual equivalence occurs at a much higher level of abstraction. Many complicated inferential procedures must go into categorising, for instance, bananas and termites as examples of chimp food. No amount of clever transforming of the retinal array is going to map **bananas** onto **termites** to create a single common representation. More abstract features such as **edibility** must come into play.

Memory of the *experienced past* is understood in terms of the persistence of stored mental representations in the appropriate neural circuits in the brain (particularly the neocortex and the hippocampal system). The *hidden present* is understood in terms of stored mental representations becoming part of the interpretation of the perception of a certain place or event. The bottom-up resources from perception fuse with the top-down resources of memory and inference to form an overall interpretation of a situation. Loftus and Palmer's (1974) research on the interaction between language and memory provides a nice example of this sort of phenomenon.

The *unexperienced past*, the *future*, and the *hypothetical* are all examples of inferential and recombinative processes based upon the internal transformation of stored mental representations into new representations of states of affairs. These are not always explicit cogitative processes limited to adult human beings. For instance, it is argued (Cheney & Seyfarth, 1990; Clark & Toribio, 1994) that monkey deception and social maneuverings imply the necessity for a sort of anticipation (future representation) and counterfactual reasoning (hypothetical representation).

The Computer as Representation Engine

Mental representations are obviously a potentially powerful explanatory mechanism for researchers interested in the time-space distancing processes associated with cognitive phenomena. The challenge for psychologists and other cognitive scientists is to come up with some way of showing how mental representations can be a respectable part of the natural order, especially that part of the natural order concerned with neurobiology. In other words researchers have to come up with some way of fitting the notion of mental

representation into an explanatory model that respects the constraints of *localised control* and *naturalism* mentioned at the beginning of the chapter. Fodor (1987) puts it this way:

> I suppose sooner or later the physicists will complete the catalogue they've been compiling of the ultimate and irreducible properties of things. When they do, ... *spin*, *charm*, and *charge* will perhaps appear on their list. But *aboutness* surely won't: intentionality simply doesn't go that deep. (p. 97)

Cognitivist cognitive science has seized upon the computer as a promising source model for understanding the brain as a mental representation engine (e.g., Fodor, 1975). Computers are physical, automatic, formal systems and formal systems are understood as systems in which symbols are manipulated 'mechanically' (i.e., without regard to the interpretations of a person). Hutchins (1995a) describes the operation of a formal system in the following way:

> [T]here is some world of phenomena, and some way to encode the phenomena as symbols. The symbols are manipulated by reference to their form only. We do not interpret the meanings of the symbols while they are being manipulated. The manipulation of the symbols results in some other symbolic expression. Finally, we may interpret a newly created string of symbols as meaning something about the world of phenomena. (pp. 359-360)

The computer, as an automatic, physically-realised, symbol manipulator, respects the constraints of naturalism and localised control: It is explainable purely in physical terms and it is an object with finite displacement. A mind, understood as a rational device, could thus be some neural version of a computer. Thus the cognitivist approach:

> takes cognition to be the operation of a special mental *computer,* located in the brain. Sensory organs deliver up to the mental computer representations of the state of the environment. The system computes a specification of an appropriate action. The body carries this action out. (Van Gelder & Port, 1995, p. 1)

Promising Properties of Computers

Cognitivist theories suggest that a computer is just the right sort of physical device for realising a mental representation system because it is a purely *physical* system that supposedly manipulates *representations* (symbols) of things *automatically* and *mechanically*. These 'representation engines' are also attractive explanatory models of cognition because their structure supports *compositionality* – the generation of molecular thoughts from cognitive atoms arranged according to a particular syntax. Compositionality is often taken to be a central aspect of human cognition (e.g., Corballis, 1991; Fodor & Pylyshyn, 1988). Computers can also be potentially programmed to be partially autonomous entities; systems that can implement ongoing input-processing-output cycles without the guidance of an external agency.

Automaticity and Partial Autonomy

The vast majority of computers in existence today require the constant attention of a human operator to provide them with input, instructions for manipulating that input

(usually in the form of pre-written software loaded onto the system) and to make sense of the output. The computer is a machine for automatically manufacturing and storing information (in the everyday sense of *information*), albeit a hugely flexible and versatile machine. Thus the typical personal computer is not a lot different from a television, car, or washing machine. One 'instructs' it to do something, lets it 'manipulate' or 'interpret' one's instructions, and then makes use of the 'output'. For example, an automatic washing machine takes an internal state given by its control settings and, via the mechanical configuration of its insides, 'transforms' this 'information' into various other types of internal states relating to water flow, temperature, spin cycle, and so on.

Computers only really begin to excite the psychological imagination when (1) this input-processing-output process is turned in on itself to become a closed loop and (2) the procedures carried out by the central processing unit are made flexible enough for the computer to, in effect, program itself. High-level artificial intelligence languages such as LISP provide this sort of potential. A computer that can sample its environment, manipulate that sampling in a meaningful way, and then output it in order to resample the environment begins to look much more like an autonomous[21] entity. That is, in principle a computer could exhibit adaptive behaviour in a 'real-world environment' for a reasonable period without the outside assistance of a beneficent engineer.

Compositionality

The computer is practically defined by the fact that its basic structure can be configured, by programming it or loading a program into it, to create a potentially infinite variety of specialised virtual machines[22]. That is, the basic 'hardwired' operations of a computer can be compiled in various combinations into an unlimited number of 'high-level' operations. Thus a computer can be, as we all know, a spreadsheet machine, a word-processing machine, a drawing machine, and so on. This fact is especially attractive for cognitive theorists for it suggests that a purely physical machine can exhibit the sort of *compositionality* that is often attributed to cognizers like ourselves (Fodor & Pylyshyn, 1988). Compositionality is usually thought to be reflected in the *productivity*,

---

[21] Boden (1996a) suggests that we judge a system to be autonomous: 1) to the extent to which responses to the environment are direct rather than indirect (mediated by internal structures shaped by the system's history); 2) the extent to which control mechanisms are self-generated rather than imposed; and 3) the extent to which internal mechanisms can be reflected upon and modified. The first two aspects are within the grasp of a closed-loop computer. It is unlikely that many cognitivist researchers would concede that existing AI systems could accomplish the third feat. In chapter 6 I discuss a much stronger notion of autonomy that requires the system in question to have a special kind of physical cohesion (Christensen & Hooker, 2000, in press). No existing computational or robotic system is autonomous in this strong sense.

[22] Actually this characterisation applies only to the ubiquitous general purpose machine.

systematicity, and *inferential coherence* of human thought and language. Productivity (or generativity) is the ability to create an infinite number of thoughts or utterances by the syntactic recursion of simple elements. Systematicity is the ability to recombine elements of cognitive expressions in meaningful ways so that, for instance, a system that can think that 'the cat is under the car' can also use those conceptual elements to think that 'the car is under the cat'[23]. Inferential coherence is the ability to make cogent inferences such as inferring "from the proposition 'It is raining in Spain' to the proposition 'It is raining somewhere.'" (Rowlands, 1999, p. 175). In order for these cognitive capacities to exist, classical cognitivist theorists (e.g., Fodor, 1975) argue that mental representations must be *symbolic*, *canonical*, *structured*, *static*, and *abstract* entities (Shanon, 1991, pp. 361-362; see also Harnad, 1990, p. 336). These are just the kinds of properties that conventional computer symbols seem to exhibit.

Symbolic representations are *symbolic* in the sense that they have two aspects: a *bearer* (or vehicle) and a *content* (or message). The bearer is of no significance except as the 'carrier' of information. The same information can be instantiated in different media (the, so-called, *multiple-realisability thesis*).

Symbolic representations are *canonical* in the sense that they are phrased in a single code. This code is *complete* (all knowledge possessed by an agent is represented by the code), *exhaustive* (there exists no subtleties of interpretation or connotation that are not explicitly encoded), and *determinate* (each distinct piece of knowledge is represented by a distinct representation that has a single interpretation).

Symbolic representations, at least as they are understood in classical computationalism, are *structured* in that they are, or are composed of, *well-defined* primitives (a finite, small set of building blocks rather like an alphabet) and *well-formed* composites (each 'word' that is created by combining groups of the primitives is done so by a system of rules – a syntax or grammar). Many connectionist theorists part company with classical computationalists on these points. Connectionists tend to view the relationship between input and output (say a question and an answer) as a complicated process that deals with 'microfeatures' or 'subsymbols' that themselves are derived from training with particular kinds of paired inputs and outputs. In connectionist networks there are no well-defined primitives. Well-formed composites (e.g., grammatically correct utterances) emerge from the connectivity

---

[23] Systematicity is very similar to Evans' (1982) notion of the *generality constraint.* Evans' idea is that we do not feel that a person, or other system, exhibits genuine thought unless they (or it) can show that they have (or it has) the capacity to juxtapose arguments and predicates. Russell (1996) expresses this idea simply when he writes "it makes no sense to attribute to somebody a thought such as Fa [where *F* is a predicate such as 'white', and where *a* is an argument, such as 'car'] unless she also has the capacity to think Fb or Fc or Fd, and so forth, as well as Ga or Ha or Ka, and so forth." (p. 165) (see also Clark, 1993).

and weightings of the network rather than by following explicit transformational rules. Arguments rage over whether connectionist networks can provide us with adequate models of cognition for these very reasons.

The 'natural state' of symbolic representations is a *static* one. Unless they are 'damaged', altered, or 'overwritten' representations are thought to be permanent. Such a view stands in opposition to dynamic models of mind which include biologically-realistic neural nets (Globus, 1992), artificial life neural networks (Parisi, 1997), and dynamical systems models of cognition (Van Gelder & Port, 1995).

Symbolic representations are *abstract* in two senses: (1) they are *amodal* or cross-modal (they have no special affinity with any particular sensory modality or way of learning; for instance, it matters not whether a representation is created after witnessing an event, hearing someone talk about it, or reading about it), and (2) they can be instantiated in many physical forms (the physical composition of the representational system is unimportant).

Simply put, mental representations conceived in these terms are thought to possess the flexibility and interchangeability necessary to create and transform cognitive expressions in characteristic human ways. It is argued, for instance, that in order for us to understand what changes when we reverse the roles of the 'cat' and the 'mat' in the sentence 'the cat is on the mat', we must possess discrete mental structures with the semantic content 'cat' and 'mat' as well as a complex group of syntactic rules for recombining them into meaningful expressions. If the 'cat on the mat' structure was not decomposable (say, it was constituted by a holistic structure) then we would not be able to perform such feats of thought.

Not all cognitive scientists are convinced that productivity, systematicity, and inferential coherence are definitive qualities of human cognition, and still fewer are convinced that a classical symbol system architecture, such as that instantiated in a typical computer, is necessary for their implementation (Rumelhart & McClelland, 1986)[24]. In addition, the concept of systematicity puzzles some (e.g., Copeland, 1993). These worries aside, there can be little doubt that the highly flexible and modifiable structure of the computer makes it a much more attractive metaphor of the human mind than one based on other artifacts.

Potential Rationality

Once the wordly event has been encoded as some sort of internal state of the machine, it can be transformed through the execution of various internal processes (often called rules or effective procedures) into a new internal state. This state can be further transformed or it can be used to create some sort of output event such as a display on a VDU, a printout, the

---

[24] For further discussion of whether connectionist networks can, or need to be able to, exhibit compositionality see Bechtel (1993a), Smolensky (1988, 1990), and chapters 7 and 8 of this work.

movement of a robotic limb, and so on. This output event may, or may not, appear to be meaningful, useful, or expected by the computer user. Such things depend on the ways in which the programmer has organised the particular virtual machine in operation. Typically the operation of a particular virtual machine is intended to 'make sense'. That is, the machine's output events are 'sensible' transformations of its input events.

Symbolic/Representational

Perhaps the most significant attraction of computers is that they seem to provide us with a purely physical way of understanding how a system can represent things in the world. Specifically we can instruct (program) a computer to recognise some sort of input (a particular key press, sound in a microphone, pattern of activation on a photovoltaic array) as relating to a certain state of the world. When an input device is activated it configures part of the internal structure of the machine in a certain way. This internal configuration is often called a symbol or representation because it correlates with a particular worldly event. Indeed, the whole point of programming a computer is so we can 'encode' wordly events as internal, storable, configurations of the machine. Not only can *we* interpret various internal states as being about things in the world, we can also begin to understand how the internal states might be representations of things in the world *for the machine.* That is, the internal states of a machine can not only be external, personal-level, public representations but also internal, subpersonal-level, mental representations. The former represent by virtue of their *derived intentionality*, while the latter represent by virtue of their *intrinsic intentionality* (Searle, 1992). It is this latter notion of internal states as representations for the machine (or mental representations) that excites the cognitivist imagination.

Of course, the trick is to argue just how an internal state is 'meaningful' for a machine rather than just being interpretable (by a clever human) as such. I imagine that few cognitivists would be game enough to argue that the internal states of a regular home computer running a word processing program were 'meaningful' for the computer (although Newell [1990] and Vera and Simon [1993] may). Cognitivist philosophers of mind tend to think that it requires a special sort of virtual machine for symbols to have intrinsic meaning. In order for a computer to be a cognitive system, some special kinds of relation must be set up between the system and the world in order for its internal configurations to be understood as mental representations. Unfortunately for cognitivism there is little in the way of consensus among modern cognitive theorists about the nature of this special set of relations. So cognitivism's first challenge is to face up to what has been called the *symbol grounding problem.*

Problem 1: The Symbol Grounding Problem

The symbol grounding problem is the problem of how subpersonal mental representations can be intrinsically meaningful (Harnad, 1990). We know that external, public representations can be meaningful because people, as interpreters, can understand and use them. But we also refer to entities within computer programs as symbols and representations. Many cognitive scientists believe that mental representations are just computer symbols realised in neural wetware instead of silicon hardware (e.g., Newell, 1990). The question is, are these computer symbols meaningful to the computer or is their meaning parasitic on a human interpreter? Harnad (1990) puts it this way:

> How can the semantic interpretation of a formal symbol system be made *intrinsic* to the system, rather than just parasitic on the meanings in our heads? How can meanings of the meaningless symbol tokens, manipulated solely on the basis or their (arbitrary) shapes, be grounded in anything but other meaningless symbols? (p. 335)

Most cognitive scientists agree that the symbols that feature in modern computer software are external, public symbols in the same way that written text is or spoken utterances are (e.g., Heil, 1981). That is, these symbols mean nothing to the computer itself. To be sure, computer symbols are automatically manipulated and transformed, but their use and interpretation is entirely up to the human user. So the challenge for the cognitivist is to formulate an account of how computer-like symbols might become meaningful. If this can be done then the cognitivist representational-computational project can be carried through.

Cognitivist theorists either believe that the symbol grounding problem can be solved or that it can be avoided. Critics of cognitivism have often raised the objection that the notion of mental representation falls foul of the *homunculus fallacy* and consequently that the symbol grounding problem cannot be solved (see Heil, 1981; Searle, 1992; Crane, 1995, pp. 149-151). Dennett (1978) describes the homunculus fallacy in terms of a paradox.

> *First,* the only psychology that could possibly succeed in explaining the complexities of human activity must posit internal representations ... *second,* nothing is intrinsically a representation of anything; something is a representation only *for* or *to* someone; any representation or system of representations thus requires at least one *user* or *interpreter* of the representation who is external to it. Any such interpreter must have a variety of psychological or intentional traits ...: it must be capable of a variety of *comprehension,* and must have beliefs and goals (so it can *use* the representation to *inform* itself and thus assist it in achieving its goals). Such an interpreter is then a sort of homunculus.
>
> Therefore, psychology *without* homunculi is impossible. But psychology *with* homunculi is doomed to circularity or infinite regress, so psychology is impossible. (Dennett, 1978, pp. 121-122)

In essence the homunculus fallacy objection claims that the explanatory terms used by cognitivism in elucidating the notion of cognition, terms such as *representation*, *computation*, *information*, *signal*, and *symbol*, are personal-level not subpersonal-level

terms and consequently that these terms are illicitly used to make explanatory mileage in cognitivist theories.

Keijzer (1997) refers to the cognitivist strategy of co-opting the personal-level conceptual framework for making sense of subpersonal processes and structures as *Agent Theory* (AT). "In AT, the subpersonal mechanism is molded after a conceptual framework that is traditionally used to describe and explain the actions of whole, behaving agents: folk psychology, or belief-desire psychology." (pp. 58-59). He makes the important point that such a strategy is not *necessarily* a problematic one. A number of philosophers and historians of science have observed that the use of a well understood process or mechanism (a source domain) to model a hidden generative mechanism thought to lie behind a puzzling phenomenon (a target domain) is an integral part of scientific explanation and theory construction (Harré, 1976; Haig, 1987). So there is no *a priori* reason why scientific explanations should not make use of models or metaphors from our everyday understandings of the world in order to illuminate our understandings of puzzling phenomena in more esoteric scientific domains. Cognitive scientists are completely within their rights to call neural states that often reliably covary with particular states of affairs in the environment *mental representations* or *cognitive states* or *green cheese* for that matter. Problems only arise when researchers claim that some sort of explanatory ground can be made by utilising the non-analogous aspects of the model or metaphor. Keijzer (1997) refers to this phenomenon as *theoretical hitchhiking*[25]. He characterises it in the following way:

> Theoretical hitchhiking occurs when the cognition-dependency of a hypothesized subpersonal property is consistently obscured by the observer's own personal-level perspective. A hitchhiking property is not really present at the subpersonal level. (Keijzer, 1997, p. 133).

Keijzer (1997) argues that we can spot a hitchhiking property when we actually try to build an artificial system using the inadequate model.

> A 'real' subpersonal property will be actually present at the subpersonal level. Its characteristics can subsequently be replicated in an artificial context and used to build an artificial, behaving system, a model. A hitchhiking property, on the other hand, provides only the illusion of being present at the subpersonal level. Models based on hitchhiking properties will reflect this incorrect analysis of the subpersonal situation that they are supposed to approximate. Crucial characteristics will be left out and the resulting model will remain flawed and not live up to expectations. (p. 134).

Sound waves, for instance, will never make someone wet. The explanatorily useful part of the wave model of sound does not carry over that part of the metaphor. If there did exist a situation where a sound of some sort was associated with someone getting wet (perhaps a particular sound is associated with sweat-inducing exercise) it would be obviously wrong

---

[25] I think that *theoretical piggybacking* might be a more apt term.

to explain the wetness 'because sound comes in waves and waves are wet'. This wave example is of course overly simple but the history of science provides us with many examples where the use of analogy in scientific explanation has foundered due to an increasing awareness of the lack of fit between source models and their targets (Laudan, 1984). In the 19th century physicists believed that light was a wave-like phenomenon. Since a wave involves the motion of some medium these physicists hypothesised that there must exist an interplanetary medium (the electromagnetic ether) through which light-waves propagated. When a number of striking observations and experiments brought the existence of ether into question the scientific view of light underwent a radical change (see Giere, 1988; Hacking, 1983). The wave model of sound illustrates the point that model and metaphor use are central aspects of theory development. The mind, no less than any other puzzling scientific phenomenon has been the target domain of many ambitious modellers.

It is important to note that the explanatory utility of any model rests upon the degree to which it is clear which parts of the model are actually doing the explaining (the positive analogies) and which bits are irrelevant and disanalogous. If it turns out that there is no useful carry over between the source and target domains then the model must be rejected. A potential explanation of a puzzling phenomenon can only proceed by using the acknowledged analogies as explanatory constructs. A theorist cannot legitimately let the often unacknowledged disanalogies do the explanatory work.

The successful use of a model from a source domain in modelling the underlying mechanisms in a target domain rests on the degree to which the model is analogous to the actual generative mechanisms at work in the target domain. In essence the homunculus fallacy objection argues that cognitivist explanations of cognitive activity rely upon a model (the computational model) that is not analogous with the mechanisms at work in cognitive activity. This can be shown by making explicit the cognitivist's tacit explanatory reliance on disanalogies in the computational model.

The Analogies and Disanalogies in Agent Theory

In terms of Peirce's understanding of representation, cognitivism latches on to the fact that representations have *content* about objects which need not be present for the representation to have an effect on the person, animal, or other cognitive agent. This provides an explanatory mechanism for behaviour in representation-hungry situations. Cognitivism also involves a clear statement that the physical *bearer* of representations in people and other animals is the brain. Although there is much argument over the appropriate *vehicle* of representations, whether they are best understood in terms of a language of thought, connectionist network, analogical, or dynamical system, most cognitivists agree that the *medium* of mental representation is the brain. So the cognitivist representational model of

the mind describes how mental representations are analogous to external representations in terms of the *bearer* and the *object* aspects of Peirce's theory of representations.

It is generally agreed however that the notion of the *interpreter* cannot be easily carried over to the subpersonal realm of the mind as the notions of bearer and object have been. This is so for the simple reason that the notion of an interpreter requires us to posit just those cognitive abilities that we are trying to explain. Thus, if we invoke an internal interpreter (a homunculus) to interpret our mental representations we will have to somehow explain how the homunculus can interpret. The only obvious way we can do this is to posit a representational system inside the homunculus which, of course, can only be made sense of in terms of another interpreter (the homunculus' homunculus).

Cognitivist Solutions to the Symbol Grounding Problem

Crane (1995) nicely sums up the theorist's dilemma at this point:

> The problem is this. Either [mental representations] get their meaning in the same way that public [external representations] do, or they get their meaning in some other way. If they get their meaning in the same way, then we seem to be stuck with a regress of homunculi. But if they get their meaning in a different way, then we need to say what that way is. Either way, we have no explanation of how [mental representations] mean anything. (p. 150)

There are basically two kinds of cognitivist response to this problem: the first is to try and argue that there is no problem, that we can possess an adequate notion of mental representation without having to introduce some system for interpreting representation bearers. The second type of solution admits that bearers must be used or interpreted in some way and that there must exist some kind(s) of natural, physical subpersonal interpreter that can perform this task. The second reply is by far the most widely endorsed by cognitive scientists.

There is No Symbol Grounding Problem

One widespread reaction to the homunculus fallacy is that it presents no problem for the cognitivist theorist because it is simply not the case that the use of representations require an interpreter with all of the powers of an interpreting agent. Dennett (1978) is probably the most well-known defender of representation via a strategy of 'stupifying homunculi'. His argument goes like this:

> Homunculi are *bogeymen* only if they duplicate *entire* the talents they are rung in to explain. ... If one can get a team or committee of relatively ignorant, narrow-minded, blind homunculi to produce the intelligent behavior of the whole, this is progress. A flow chart is typically the organizational chart of a committee of homunculi (investigators, librarians, accountants, executives); each box specifies a homunculus by prescribing a function *without saying how it is accomplished* (one says, in effect: put a little man in there to do the job). If we then look closer at the individual boxes we see that the function of each is accomplished by subdividing it via another flow chart into still smaller, more stupid homunculi. Eventually this nesting of boxes within boxes lands you with homunculi so stupid (all they have to do is remember whether to say yes

or no when asked) that they can be, as one says, "replaced by a machine.". One *discharges* fancy homunculi from one's scheme by organizing armies of idiots to do the work. (Dennett, 1978, pp. 123-124)

Ultimately, what Dennett is saying here is that homunculi are merely explanatory promissory notes of potential, future explanations at a purely mechanical and structural level. We leave a limited 'mentality' in our proposed explanatory mechanisms and gradually attempt to whittle it away as we increase our knowledge. In other words homunculi are not interpreters but rather methodological artifacts. They help us do our research but we never think of them as ontologically real. The problem here, as Dennett acknowledges, is that cognitivist theorists become the unlucky possessors of an unpaid explanatory loan. That is, the concept of mental representation is used without an agreed upon understanding of what a mental representation is. We await the day when we understand behaviour and cognition in terms of a purely physical theory. In fact that is precisely what other philosophers are trying to do when they propose their theories of content. Only then will we see how to reduce mental representation without resorting to the use of homunculi. Thus the homuncular functionalist approach roughly says "Just wait and see, we're sure that we'll ultimately pull this off".

Theories of Representation

Philosophers, and to a much lesser extent other cognitive scientists, have long theorised about how representations might be meaningful. By and large the general attempt has been to formulate theories of content; that is, theories about how mental representations get their content. Call this the *problem of content assignment*[26]. This task generally *assumes* that mental representations exist somehow in mind/brains. Chemero (1998a) argues that recent arguments in favour of anti-representationalism have shifted philosophical interest to the question of what makes something a mental representation in the first place. We might call this the *representation constitution problem*[27]. Clearly the latter problem is more central to questions about the theoretical status of cognitivism. However it is important to realise that this distinction is not always made and that many theories of content tell us something about what it is for something to be a representation (see, e.g., Von Eckardt, 1993). With this in mind the approach I will take here is to use insights from content theories in combination with claims about what constitutes a representation. I will discuss three basic kinds of theory of representation: indicator or causal theories, simulation or functional role theories, and teleological theories.

---

[26] Cummins (1989, 1996) calls this, or something very like it, the *problem of meaning assignment.*

[27] Cummins (1989, 1996) uses the term the *problem of meaningfulness.*

Representations as Indicators: Causal Theories of Content

Causal theories, also known as indicator theories and covariance theories, are currently the most popular kind of theory of content (summaries of these positions can be found in Crane, 1995; Cummins, 1989, 1996; Rey, 1997; Von Eckardt, 1993). The basic idea of causal theories is that internal representations are about what causes them. Unfortunately such a crude idea is obviously far from adequate. The first problem is that we know that not all 'thoughts' about something are caused by being exposed to their object. Some thoughts about sheep, for instance, are caused by thinking about farm animals in general, others by mistaking a goat for a sheep. Indeed this fact seems to be central to the cognitivist idea that representations enable complex behaviour by standing in for things that are not currently perceivable. The second problem is that not all encounters with an object cause us to produce a 'thought' about that object. Sometimes sheep are mistaken for goats and thus a goat-thought is also caused by sheep.

So causal theorists require a more subtle criterion of content determination than that given by a crude causal theory. Instead they appeal to the notion of *reliable indication.* This is just to say that a mental representation represents an object only when the object is actually present, or is really being thought about, or conditions are such that it would be reasonable to believe that object is present. The problem then becomes one of explaining how, and under what circumstances, a bearer reliably indicates its object. Some, like Dretske (1981, 1988, 1995) argue that historical factors (evolution, learning) are responsible for the emergence of reliable indication. Others, such as Fodor (1987, 1990), claim that there exist logical factors which determine whether something is a representation. However, even if such a definition of reliable indication is possible, causal theories face a third significant problem; that it is by no means clear how a causal theory can account for the possession of representations of things that cannot be perceived – things such as Santa Claus, mathematical functions, and future events. Generally causal theorists have had their hands full trying to make sense of representations of things that do physically exist, so this problem has not been broached. Perhaps, argue the causalists, these more complex mental representations are the result of some further process that derives them from basic perceptually-based representations. If we can make sense of these basic representations *then* we can move on to the more complicated examples. Until then we should try and make sense of simpler causal representations. However, these cases seem to raise problems of their own that few philosophers believe have been satisfactorily solved.

The key problems facing indicator-type causal theories revolve around the much discussed *Problem of Error.* Simply put the problem is that indicator explanations do not explain how it is possible to misrepresent things. Most philosophers argue that, if anything is characteristic of a representation, it is the fact that a representation-user can *misrepresent* events and objects. Misrepresentation is important for cognitivism because it relates to the

central assumption that representations can stand in for objects in their absence. A system that misrepresents can obviously do this. The problem of error can be understood by thinking of it in terms of two related subproblems: the *problem of misrepresentation* and the *disjunction problem.*

The misrepresentation problem states that since representations, understood as reliable indicators, will always indicate *something* it is hard to see how they can ever misrepresent an object by triggering the *wrong* mental representation. If, for instance, we take a certain type of spot to be a reliable indicator of measles (i.e., the spots naturally represent measles) it does not make sense to suggest that the spots could misrepresent something else. *People* may take the spots to indicate, say small pox, but there is no way that the spots could *naturally* misrepresent the disease. A natural sign of something always reliably indicates its object. We know, however, that our mental representations (thoughts, plans, memories) can often be in error. So indicator theories do not provide an adequate account of this aspect of mental representation.

Dretske (1988) tries to solve this problem by suggesting that misrepresentation occurs when an organism *uses* a natural representation bearer and the bearer is incorrectly configured or activated due to a system malfunction or less than ideal conditions. As we have seen, he refers to these representations as *type III representations*. For instance, as noted above, Dretske (1986) argues that water-born bacteria containing magnetosomes (subcellular components that point to magnetic north and are 'used' by the bacterium to orient itself) constitute a natural representation system of 'safe water' (because their pointing north directs them away from toxic, oxygenated surface water) and that this system can misrepresent when a magnet is placed near the bacterium.

Dretske's solution has been criticised in two of ways. Cummins (1989) argues that this malfunction view of misrepresentation does not account for the fact that not all misrepresentation occurs because of problems in the representation-constructing and representation-using processes. Sometimes we intentionally misrepresent things (e.g., when a child pretends that a big box is a house). More recently he has argued that we should distinguish between *error* and *falsehood* (Cummins, 1996). Error occurs when the representational content is inappropriate to the situation, whereas falsehood occurs when an object is represented as something that it is not. We may have a false representation of an object that is none-the-less a correct use in a particular situation. Thus Cummins' analysis demands an understanding of mental representation that can deal with both error and falsehood. Because Dretske's theory treats all errors as falsehoods (and vice versa) it cannot answer this criticism. It may be possible to modify Dretske's theory to handle this problem, but the problem itself indicates that an adequate notion of misrepresentation, and thus representation, is more complicated than is generally supposed.

A possibly more serious problem for Dretske's theory becomes evident when we try to specify the content of the natural representation in question. Take the magnetotactic bacteria example mentioned earlier. Hendriks-Jansen (1996) argues that this is not a simple or obvious example of misrepresentation because there exists no simple way of determining just what it is that the magnetosome system actually represents (if it represents anything). It may represent toxic water, danger, north, magnetic north, geomagnetic north, or even the presence of a relatively strong magnetic field. The simplest 'object' of this system would be the last mentioned. But if that is the case, then a nearby magnet does not cause any misrepresentation.

Dretske is not unaware of this sort of problem. He suggests that we can better identify the content of a representation when information from multiple sensory channels is causally connected to a single bearer (see also Lloyd [1989] for a similar point of view). In such cases where, for example, multiple proximal stimuli such as the sound, sight, and smell of an approaching car, can all independently trigger a single internal structure (a representation bearer), we are in a much better position to argue that the structure represents an approaching car (a distal stimulus). In situations, such as the magnetotactic bacteria, where there exists only a single information channel, it is impossible to conclusively argue that an internal structure (bearer) 'represents' a distant environmental object, event, or condition (e.g., toxic water), rather than a simple local stimulus (e.g., the presence of a magnetic field). The existence of multiple triggering conditions is consistent with, but not sufficient for, concluding that the bearer is 'causally connected to' the environmentally significant object. If this sort of analysis is plausible, then indicator theorists are in a better position to argue for a natural, causal account of misrepresentation because it seems much easier to appreciate cases of misrepresentation when dealing with distal stimuli (e.g., toxic water, approaching cars) than proximal stimuli (e.g., the presence of a magnetic field, a rapidly increasing roaring sound).

Unfortunately the lack of sufficiency here plagues this sort explanation. The mere fact that information from multiple sources triggers a single bearer is not enough to conclude that it is connected to a single distal stimulus rather than a collection of proximal stimuli. And if this is the case, then it becomes problematic knowing how to interpret the triggering of a bearer by multiple stimuli. How do we decide on the content of a bearer that is (apparently) mistakenly triggered by a stimulus unrelated to the distal object? Is the situation one of misrepresentation (e.g., **car** is triggered by the sound of the wind) or one of accurate representation where the content of representation is *either* **car** *or* **wind that sounds like a car**? This difficulty is known as the *disjunction problem* and forms the second important part of the problem of error. The disjunction problem can be described schematically in the following way:

If a mental representation (MR1) reliably indicates an object (O1) but MR1 is triggered by O2 on a particular occasion, then it is not legitimate to argue that O2 has been misrepresented by MR1. What prevents us from arguing that MR1 is a representation of O1 *or* O2-in-certain-circumstances?

Fodor (1990) attempts to avoid this problem by appealing to something called *asymmetric dependence*. This is basically the idea that misrepresentations are asymmetrically dependent on accurate representations. In schematic form Fodor argues that:

If it were not true that O1s cause MR1 then O2s would not cause MR1 (O2s cause MR1 only if O1s cause MR1).

and

If it were not true that O2s caused MR1 then O1s would still cause MR1 (*it is not the case* that O1s cause MR1 only if O2s cause MR1).

The main problem with Fodor's answer is, as Cummins (1989) and others have argued, that Fodor seems to be merely redescribing what is the case in a disjunctive situation rather than saying how a representational system may deal with it. Thus, Crane (1995) writes that he is

> unable to see how asymmetric dependence goes any way towards *explaining* mental representation. I think that the conditions Fodor describes probably are true of mental representations. But I do not see how this helps us to understand how mental representation actually works. In effect, Fodor is saying: error is parasitic on true belief. But it's hard not to object that this is just what we knew already. The question rather is: *what* is error? Until we can give some account of error, then it does not really help us to say that it is parasitic on true belief. (pp. 180-181)

Dretske (1986) approaches the disjunction problem in a different way. He argues that genuine misrepresentation, and thus, presumably genuine mental natural representation, can only occur in organisms or systems of a particular level of complexity (see also Lloyd, 1989). In particular an organism must be capable of simple associative (respondent) learning.

A representation-using organism must possess internal structures (representation bearers, RB) that covary with particular stimuli (S) (for simplicities sake we can gloss stimuli as signs of some object or referent O). An RB then leads to the appropriate sort of behavioural response (R). Furthermore each RB must be connected to each O by more than one S by, for example, detecting an object using different sensory modalities or by responding to different aspects of the object accessible via a single modality (e.g., its colour and style of movement). This condition goes some way toward avoiding the idea that the RB is just an indicator of a particular proximal stimulus. Of course this setup remains vulnerable to the disjunction problem. That is, it cannot tell the difference between 'sensitivity to O' versus 'sensitivity to S1 in some conditions and S2 in some other conditions'. So, in order to

avoid this problem Dretske argues that a representation-containing organism must be also capable of associative learning; of learning how to take a new S as a sign, or conditioned stimulus, of O, and thus use that new S as information for initiating B.

How can such a set up unambiguously account for misrepresentation? Dretske argues that if an animal with all of these capacities acts with behaviour B in response to CS_i (a conditioned stimulus which is normally connected to O1) but on this occasion CS_i is produced by O2 then we can legitimately suppose that the organism is misrepresenting CS_i as O1. Why? Because the animal has a history where CS_i was learned to be paired with S_i which produced B. We know that CS_i was learnt as a substitute for S_i. Dretske applies the same kind of logic where the historical factors are evolutionary rather than ontogenetic.

Does Dretske's approach adequately deal with the notion of misrepresentation? In a sense Dretske uses an historical explanation to prop up an indicator explanation where the historical explanation seems to require an *a priori* indicator explanation! This raises the question of how we work out what the original unconditioned stimulus is and thus what the 'true' content of the representation is[28]. Without this information Dretske's solution to the disjunction problem cannot get off the ground. In a sense Dretske assumes that we already *know* the content/object of the representation *before* he introduces the complexities of multi-modality and associative learning. Obviously such a strategy cannot be used to work out whether a system is representational or not, let alone what the content of the representation is.

Current causal theories seem to struggle with the basic problem of providing a naturalistic explanation of how an internal and subpersonal component of a cognizer can possess representational content. I think it is possible that a future causal theory of representation may eventually emerge that can explain how a system can exhibit some form of appropriate object constancy – that is, be able to distinguish various objects and events in its environment in a reliable manner. This will mean that there may exist a component of the system's innards that can act as a reliable indicator of a particular referent. Also it is no doubt possible that this reliable indicator theory will be able show how this indicator can be inappropriately triggered and thus show that the system has 'misperceived' something in its local environment. But such a theory is a long way from providing anything like the kind of understanding of representation that cognitivism requires. Indeed a reliable indicator theory does not really seem to be a theory of representation at all. After

[28] One implication of this is that Dretske's theory violates the intuition that it should be the present structure of the system, not its history, that determines whether something is a representation (see the following section on teleological theories for more on this criticism).

examining Lloyd's (1989) indicator theory of representation, and the robotic example he uses to illustrate it, Keijzer (1997) remarks that the robot:

> on the other hand, is a device which, many [e.g., J. A. Anderson, 1991; Fodor, 1986] would hold, does not depend on representation at all. Its behavior can be fully described in terms of reactions to configurations of external stimuli. From a theoretical perspective there is no need to invoke the concept of representation in this case. (Keijzer, 1997, p. 102)

Reliable indicators are really just a kind of subsystem that can be activated in response to some kind of, typically high-level, invariant (to use Gibson's [1979/1986] terminology). They are not internal simulators of the outside world but rather just sophisticated kinds of *detectors* that may be able to fix onto distal targets. I suspect that this is why causal theories have so much difficulty in dealing with misrepresentation because misrepresentation is bound up, in some sense, with the ability to think about or interpret a situation in a decoupled and off-line manner. Yet reliable indicators are subpersonal on-line instruments. They are not, it seems, the kinds of entities that can serve as cognitivism's representations. Bem and Keijzer (1996) suggest that reliable indicators may form some part of a theory of cognition but it will not be one concerned with representation.

> What we can do is to construct a theory about the indicating functions of neural elements (networks) and then, remaining loyal to this level of explanation, avoid semantics and intentionality. But *we cannot have a theory of meaning while staying within the mere indicating system* (p. 462, emphasis added).

So it seems that cognitivism requires a theory of representation where representations can modulate behaviour in the absence of environmental information – this is the core notion of functional role theories of content.

Representations as Simulations: Functional Role Theories of Content

Keijzer (1997) argues, as I have above, that representation is postulated as an explanatory concept within cognitivism in order to explain how a cognitive system can coordinate its behaviour with respect to things that the system is not able to sense. A representation, in this sense, is a model of the world that can be used in the absence of the things that it represents. One commonly discussed idea, that derives from Haugeland (1991), is that something is a representation if it can 'stand in for' for its object (see also Bechtel, 1998; Clark, 1997, chap. 8; Chemero, 1998a; Grush, 1997). Clark (1997) provides a concise summary of Haugeland's position. For Haugeland, a system is a mental representational system just in case:

> (1) It must coordinate its behaviors with environmental features that are not always "reliably present to the system."
>
> (2) It copes with such cases by having something else (in place of signal directly received from the environment) "stand in" and guide behavior in its stead.
>
> (3) That "something else" is part of a more general representational scheme that allows standing in to occur systematically and allows for a variety of related representational states. (Clark, 1997, p. 144).

The first point distinguishes representational systems from ones that only use information from the local environment to modulate their behaviour, such as plants tracking the sun with their leaves. In other words, Haugeland (1991) is arguing here that the reliable indicators of causal theories of content are not representations. Grush (1997) makes a similar claim. The second point identifies how we might begin to locate the representation bearer. The third point qualifies the second by pointing out that not all physical components that seem to stand between an event and an activity are representational. Rather, representations exist only in systems that must deal with a reasonable variety of event-activity mappings. Haugeland (1991) illustrates why this third condition is necessary by suggesting that "triggered gastric juices might keep a primitive predator on the prowl, even when it momentarily loses a scent – thus standing in for the scent." (p. 62). Haugeland (1991) seems to view such a 'stand in' as insufficiently representational because of its lack of flexibility and specificity. It does not seem to be tied to a particular environmental event in a strong enough manner to constitute a representation. Haugeland's framework is not meant to be at all rigorous, but it serves to illustrate the idea that a representation is a kind of internal surrogate for distant environmental objects and therefore useful as an internal controller of behaviour[29]. Haugeland's framework thus relies on the idea that representations are essentially internal simulations or isomorphisms of (possible) external events. This is the core idea of *functional role theories of representation* (a.k.a. inferential role theories, conceptual role theories; see, e.g., Von Eckardt, 1993)[30]. One of

---

[29] Chemero (1998a, 1998b) critiques Haugeland's notion of representations as decouplable, environmental stand-ins at length. He argues that the notion of 'reliable presence' is too vague to do any real explanatory work and that it is revisionist in nature.

[30] According to Von Eckardt (1993) functional role theories include *conceptual role semantics* (e.g., Loar, 1981), Cummin's (1989, 1996) *interpretational semantics*, and Newell's (1990) approach to representation. In the conceptual role semantics view a representational system is a computational system whose relations (transformations, associations, implications, entailments and so on; these are sometimes referred to as *epistemic liaisons*) between bearers mirror the relations that hold between entities in the world or entities in logical space (see also Lloyd, 1989, pp. 24-37). So representations get their meaning primarily in terms of the ways in which they relate to other bearers. This makes intuitive sense in that, for instance, we can imagine learning the meaning of a new word without having its referent pointed out because we come to understand how it relates to other words (representations) that we already know. Criticisms of conceptual role semantics usually focus on the fact that they do not appear to say anything about the external referential relationship concepts have with things in the world. *Dual aspect theories* (e.g., Block, 1986) attempt to solve these problems by thinking of meaning as a combination of internal relations between other concepts and external relations between concepts and objects in the world. For the purposes of this thesis I will consider simulation representations to be of the dual aspect variety rather than purely internalist in nature. This, I believe, seems to be what theorists such as Cummins and Newell have in mind in their formulations of representational content.

the clearest expositions of the functional role approach is the interpretational semantics approach of Robert Cummins (1989, 1996).

In his approach Cummins (1989, 1996) attempts to reformulate the notion of mental representation in order avoid the problems implied by theories of representation that require recourse to an interpreter. He argues that all 'use theories' of representation are doomed to failure because they do not adequately characterise what it is about representations that makes them represent. In essence he argues that notions of 'use' and 'interpretation' are red herrings for a theory of natural representation. Natural representations represent, so he argues, because they are isomorphic to the intended referent. For Cummins a natural representation represents its object regardless of whether or not it is used and regardless of whether or not it is used correctly. The essence of the representational relationship is the preservation of isomorphic relations between the bearer and its object. To capture this idea he refers to such entities as *simulation representations* or *s-representations*. Thus, Cummins distinguishes the question of the determination of representational content from that of representational use or interpretation.

He further argues that symbols form only an unusual subset of representations and make poor explanatory entities for theories of cognition. A symbol, says Cummins, is the sort of representation that *does* require 'use' or 'interpretation' to have content assigned to it. The bearer of a symbol has an arbitrary shape or structure. That is, it is not in any way isomorphic with its object. It gets its content by being grounded in natural representations that are isomorphic with their objects. For instance, a stop sign (a symbol) gains its meaning by being parasitic on social conventions realised by human minds that possess mental representations (natural representations of the isomorphic variety). Cummins closes his argument by noting that isomorphic natural representations do not need to be grounded (have elaborate causal or conceptual connections with their objects) because the isomorphism relationship is sufficient for characterising the representational relationship. This implies that there is no symbol grounding problem. He argues that:

> for symbols, being meaningful just reduces to being understood, and ... *understanding* a symbol (or any kind of representation) requires relating to something else. Symbols don't mean, in other words, except to the extent that they meanfor. But for nonsymbolic representations like maps, being meaningful does not reduce to, or even require, being understood or having a meaningfor [i.e., being used]. Maps represent whether or not anyone does or can understand them. ... [T]he content of a map is independent of its "grounding." ... Grounding is therefore not an issue for representation, but for understanding, and relates to meaningfulness only because symbolic meaningfulness requires understanding, that is, because symbolic meaning is based in meaningfor, and is therefore not really a *semantic* matter at all, but a cognitive one. (Cummins, 1996, p. 130)

Cummins raises the important point that there is a sense in which the structure of a bearer (of a nonsymbolic representation) *does* 'contain' information. There has been a tendency among many cognitive theorists, who are used to thinking of representations as symbol-like, to think that a bearer has a completely arbitrary connection with its object and thus

that the burden of making use of the representation is mainly the province of the interpreting system[31]. But Cummins' useful corrective shows us that the structure of a bearer limits the possible informational interpretations that can be derived from the use of the bearer. However, despite his insights there seem to be two crucial difficulties with the idea that representations are simulations.

The first is that Cummins seems to ignore the fact that the structure of a representation bearer is highly likely to be consistent with multiple, perhaps infinite, different interpretations of what the bearer may represent. It has long been argued that the entities and relations that hold in a model are isomorphic with a large number of worldly event referents or natural functions[32]. If this is the case then cognitive scientists face a massive problem trying to determine the content of a representation. Since simulation representations can be used in the absence of their referent, unlike reliable indicators, there is no way we can even get a clue as to whether a particular representational transformation is, to use Pinker's (1997, pp. 80-81) example, a case of chess playing or a re-enactment of the Six Day War.

But things get worse. Searle (1992) and others have argued that *any* physical system could be interpreted as computing some kind of natural function. He argues that the atoms in his office wall may be actually computing the same functions as his computer's word-processing software. Interestingly defenders of the computational view seem to agree with Searle, but constrain things by arguing that cognition is the only natural phenomenon in which computations are *constitutive* of that phenomenon (Chalmers, 2000). The problem then becomes one of knowing when 'computation for cognition' is occurring and when the computations are merely a side show to some other natural phenomenon such as digestion or being an office wall. So, not only does the isomorphism view of representation seem to

---

[31] Indeed Newell's (1990) notion that the defining property of a physical symbol is its ability to encourage *distal access* is entirely in keeping with this symbol-based view. The idea here is that a symbol serves as a 'pointer' to meaning or knowledge stored elsewhere. In this case the symbol does not so much represent, as serve, an information-fetching role. In this case the symbol only needs to be recognisable by the system that uses it and can take pretty much any form at all. It becomes a single-valued entity that points to a particular location rather than an information bearer. In sum a physical symbol contains little information itself.

[32] Winograd and Flores (1986) provide a nice illustration of the multiple isomorphism problem when they discuss how 'display hack' programs use the complex internal 'representational' structures in computer programs to produce geometrical designs on a video display: "Many of these grew out of programs that were originally created to perform operations on symbols that had nothing to do with visual figures. When the contents of some of their internal storage cells were interpreted as numbers representing points on a video display screen, strikingly regular patterns emerged, which the programmer had not anticipated. ... In these cases, the description of the program as representing something is a description by some observer after the fact, rather than by its designer." (p. 92)

preclude finding out the content of a representation, it also seems to make spotting a representation well nigh impossible.

One could argue that such examples do not make simulation theories problematic because a representation represents what it was originally *designed* to represent not what it accidentally causes when it is plugged into a different sort of interpretational system. But such a response just reinforces the point that Cummins goes to some lengths to avoid, that the structure of the bearer alone does not specify a unique representational content. In order to do this we need to take into account the ways in which the bearer is used and perhaps even the purposes for which it was designed (see below for the problems these extra content-specifying criteria raise).

It seems then that everything is a simulation representation of something or multiple somethings. The fact that a bearer cannot represent just anything (i.e., be completely arbitrary) does not mean that its content is unique. Cummins seems to shift all of the difficulties associated with explaining natural representations to the 'cognitive' or understanding processes that must be part of the cognitivist story. For instance, it now becomes a burden of 'cognitive' processes to work out which isomorphism is the right one for a particular representational relation – that is of learning to apply the appropriate *interpretation function* (Cummins, 1989, 1996).

However, these criticisms can be stayed to a certain extent by the fact that robotics research has shown that systems can be constructed that use simulation representations (or at least internal physical structures that are isomorphic to structures in their environments) to control their behaviour. Thus, Keijzer (1997) argues that "[t]he empirical context imposes constraints that limit the open-endedness of s-representations ...", because within robotic simulations these "representations are used as a causal factor that helps to explain how a behaving organism coordinates its activity with respect to distant and future events in the environment. The behavioral context restricts the number of systems to which a representational interpretation can be applied, and it also limits the number of possible interpretations." (p. 95). However, it is far from clear that such systems can actually engage in something akin to natural cognitive activity (I will discuss some of the problems with model-based robotic architectures in the next chapter). So it is still by no means clear that s-representations are plausible candidate internal mechanisms for natural cognition. It is also unclear how we could tell whether a natural cognizer possessed s-representations.

However, there is another avenue open to the representational theorist. They can argue that natural selection would see to it that the appropriate relations exist between the world, a representational system, and behaviour and bypass the question of the kinds of causal structures that realise these relationships. These kinds of arguments are central to teleological theories of representation.

Representations as Teleological Entities

A number of theorists have argued that we can only understand representations relative to an animal's history of natural selection (e.g., Chemero, 1998a; Dretske, 1988; Millikan, 1984, 1993; for summaries see Crane, 1995; Hendriks-Jansen, 1996; Lyons, 1992; Von Eckardt. 1993). The basic idea here is that a representation is a correlation between an inner physical structure (a bearer) and a state of the environment (an object) that has the *function* of adapting the organism, by taking part in the production of appropriate behaviour, to that state of the environment. The reason that the representation has this function is that it has been selected to do that job in the evolution of the organism. Under this scheme representation systems are viewed in the same way as other phenotypic traits – as things that have been selected because of their functional advantages to the animal. Thus representation systems have evolved because at one time or another they contributed to the fitness of the animal. This basic picture is complicated somewhat when one pays close attention to the logic of natural selection.

Typically phenotypic traits are selected because they enable a large enough members of that species to survive and reproduce for the lineage to continue. This does not mean that the trait will bring success in all or even a majority of cases. For instance, the function of acorns is to become oak trees even though most acorns will rot. But the reason that acorns exist now, with the properties they have, and occurring in the numbers that they do, is that they have enabled the continuity of the oak tree lineage. Millikan (1984, 1993) refers to the conditions under which a trait was selected as *normal conditions*, even though normal conditions may only obtain infrequently. Phenotypic traits thus have what she calls a *proper function*, a term that refers to the job that they were selected to do even if they rarely or never actually manage to achieve that job (more on this term below).

Millikan uses this basic framework to develop an explanation of what she calls *intentional icons*. Mental representations are one kind of intentional icon. Chemero (1998a) provides us with a stripped down version of Millikan's theory that preserves all of its essentials. The basic idea is that a representation is part of a naturally selected biological representation system. This system stands between a biological representation producer and a biological representation consumer. A producer produces a representation (e.g., a perceptual system, a speaker) and a consumer uses the representation (e.g., a motor system, a person comprehending speech). The proper function of the representation is to adapt the consumer to some aspect of the environment by enabling the animal to behave appropriately with respect to that aspect of the environment. The key term in this sentence is *proper function*. A proper function is the function something was 'designed' (in this case by evolution) to perform. So a faulty heart, for instance, may still have the *proper* function of pumping enough blood about the body for survival even if it currently cannot achieve this task. The content of a representation in this view is the way the world would need to be (the normal

conditions) for the representation to achieve its proper function. Thus, a representation can be said to misrepresent when it functions under non-normal conditions. Chemero (1998a) describes it in the following way:

> Since the content of a representation is determined by its function, along with those of the representation producer and consumer, its content will remain constant even in cases in which one or more of the producer, consumer and representation itself fails to work properly. That is, just as the function of a sperm is to fertilize an egg, despite the fact that the number that do so is vanishingly small, so the function of an icon that represents "chicken-here-now" does not change, even in cases in which the icon is produced or used improperly. (p. 27)

Roughly speaking, Millikan's use of natural selection seems to provide her with a scientifically sound way to argue that misrepresentation occurs when a representation does not represent what it is supposed to.

Millikan's teleological theory has some fairly radical implications that she is fully prepared to accept. One important implication is that it is not possible to tell whether a particular structure *is* a representation (i.e., solve the problem of representation constitution) or what its content is (i.e., solve the problem of content determination) just by examining its physical properties. This is because a teleological approach argues that the status of representation is historically determined, not causally determined. Indeed Millikan argues that of two 'internally identical' physical structures that possess the same relations to other things in the world ('externally' identical) only one may really be a representation with representational content. Lyons (1992) provides an apt analogy. He suggests that a representation might be akin to a key. An object is a key only if it was designed to be a key, and only a key of a certain lock if it was designed for that purpose. The fact that another (undesigned) object may function as a key, or the key functions to unlock other doors, is just good luck – a fitting of a structure to part of the world for which it was never designed. The piece of metal that fortuitously opens a lock is not a key for that lock, and a key designed to open a lock that nevertheless opens it only on rare occasions is (always) a key for that lock.

Such radical thinking has met with a host of criticisms[33]. The first involves a suspicion of the teleological implication that only evolved creatures can have representational systems. Millikan does not see this as a problem, but the intuition among many theorists is that a spontaneously emergent molecule-for-molecule duplicate of an evolved agent (the infamous 'swamp man') would possess all of the powers (cognitive and otherwise) of the

---

[33] I think that even if Millikan's theory turns out to be plausible, its portrayal of mental representation is so radically unlike the loose conception that many psychologists have of the concept, that it would not really solve *their* problems of representation and cognitive order (see Hendriks-Jansen, 1996 for an argument in favour of much of what Millikan says except for her attempt to naturalise mental representation).

evolved version of the agent[34]. It is hard to see how an historical explanation can make sense of the traditional notion of mental representation as some sort of bodily structure. Cummins (1996) argues that the problem lies in the fact that teleological theorists try to explain why a structure is a natural representation by reference to natural selection rather than explaining why a structure is selected by reference to its representational qualities:

Adaptational stories get the explanatory order wrong. Representations, when they are adaptive, are adaptive because they represent what they do. Think of cognitive maps (Tolman, 1948): They are adaptive because they are isomorphic to the spaces they map; they are not isomorphic to the spaces they map because they are adaptive. Explaining correctness in terms of adaptation gets matters backward. The adaptational theory is motivated by the fact that it is plausible ... to suppose that correct uses of representation are adaptive. (Cummins, 1996, p. 46)

The second common criticism is that teleological theories only provide us with an explanation of the system's general adaptational workings, not the adaptive particular instances of representation use. That is, these theories seem to explain the existence and usefulness of *systems* or *mechanisms* for constructing representations rather than the individual representations themselves (see Fodor's Donellan Lectures cited in Lyons, 1992; Sterelny, 1990). Lyons (1992) expresses this criticism in the following way:

[T]o appeal to evolution as a way of individuating mental content is a very crude instrument. Evolution is the 'selection' of *types* of organism by the environment on the basis of behaviour. So, at most, what can be said to have survival value is types of behaviour not particular refined or precise instances of mental content. (p. 320)

Lyons (1992, p. 319) suggests a possible retort for the teleological theorists: they can note that one just needs to find out about the situation that the agent finds themself in and 'apply' the evolved structures and processes to it to derive the actual content of the representation in question.

A third problem for teleological theories relates to the explanation of modern human mental states and representations. There obviously cannot be a simple evolutionary explanation of non-survival relevant representational states such as the desire **to go bungy jumping**. This complaint is really an implication of the previous criticism, but it involves a further twist – the idea that many of our knowledge structures seem to be products of

---

[34] For those who may baulk at the swamp man thought experiment because it asks us to imagine an impossible event, Snowdon (1988) raises a related but more realistic problem: "Due to a random mutation, a creature is produced with a new sense organ and connected neural system which provides it with information about aspects of the environment. The system performs a highly advantageous function and persists and spreads. It therefore gets selected to perform that function. However, how do we characterize in cognitive, representational terms the very first creature to possess it? In it, the system had not been selected for anything, because the processes of selection had not [a] chance to operate." (p. 630). Eliasmith (2000) provides a similarly plausible explanation in terms of connectionist networks.

cultural, social, and individual processes rather than being evolutionarily determined. This suggests that teleological theories are hugely inadequate as a means of 'reading' most of what is going on in human minds.

Teleological theories have also been criticised for their seeming reliance on an *optimality view* of evolution where every trait (in this case every representational system) is selected for its fitness-enhancing (in this case accuracy of the representations generated) qualities, when in fact we know that not all traits are fitness relevant (some are neutral and some may even be unfavourable but piggy-back upon the genes associated with other important traits, Hendriks-Jansen, 1996). In a review of Millikan's theory Godfrey-Smith (1988) remarks that

> Her model explanations are biological, yet it is well known that within biology there has lately been a reaction against the 'Panglossian' or 'optimising' view of natural selection. Stephen Jay Gould and Richard Lewontin spearhead the reaction, claiming that the products of evolution are not collections of independently selected, perfectly functioning parts, but a tangle of engineering compromises, including many features with no-function. (p. 559)

The introduction of concepts such as *exaptation* and *spandrel* into the adaptational argument (Gould, 1991) suggests that it is possible that mental representations and representational systems may, in some way, have no proper function and thus have no content. A number of evolutionary psychologists (e.g., Barkow, Cosmides, & Tooby, 1992; Dennett, 1995; Pinker, 1997) have argued that spandrels and exaptations pose no major threat to an optimality view of evolution. Spandrels are accidental byproducts of evolved phenotypic structures used by intelligent creatures for a certain non-evolutionary purpose (e.g., the bridge of the nose being useful for holding up one's glasses). Such features do not evolve because of their teleological functions. Evolutionary psychologists argue that representational systems are obviously not spandrels and therefore the force of the spandrel argument is discharged.

Exaptations are phenotypic traits that have acquired a new function that either co-exists with an old function or replaces it. Natural selection cannot account for the *origin* of the new function but it can account for the preservation of that function in subsequent generations of a lineage. In other words, as soon an exaptation is maintained because of its evolutionary benefits it becomes an adaptation. Thus, exaptations seem to be relatively minor, short-lived aspects of the evolutionary process and thus not do not pose a major problem for a broadly teleological analysis of mental representation. Millikan (1993, pp. 41-50), however, *does* seem to consider the exaptation argument to be an important one and has modified her basic theory to take account of them. After all, arguing that exaptations are exceptions to the general rule of adaptation does not discharge the argument that there may well exist some adaptive 'representational' structures that have no adaptational history. Thus Millikan (1993) argues

Being 'built' by natural selection is sufficient for proper function, being maintained by natural selection is independently sufficient, and having been utilized by other structures built or maintained by natural selection is also independently sufficient for proper function. It is hard to see how modern human cognition could fail to be caught in one or more of these nets. (p. 49).

These difficulties with making sense of the teleological functions of particular body systems may reflect a deeper difficulty with the use of the modern Darwinian-Mendelian synthesis in our explanations of mental representation. Representations and representation-systems are widely held to be discrete *structural* traits of an organism, whereas evolutionary analyses often focus upon the evolution of *whole* living phenotypes – after all, it is organisms that evolve not specific parts of them[35] (Hendriks-Jansen, 1996; Oyama, 1985). Although it is not totally incorrect to say that organs evolve under selection pressures, it needs to be clearly understood that they evolve only indirectly in terms of the behaviour that they enable. The fitness enhancing benefits of any bodily structure can thus only be understood in relation to the rest of the organism's body and the ways in which that body enables behaviours within a particular ecological niche. In this sense function (and content) of a body structure is highly reliant upon other body structures as well as the characteristics of the niche in which that body structure evolved.

Teleological theories attempt to bring the developed theoretical apparatus of evolutionary biology to bear upon the issue of mental representation in the explanation of behaviour. Like Hendriks-Jansen (1996), I think that this is an important development but one that, taken to its logical conclusion, would reject the notion of mental representation altogether. Teleological theories have difficulty reconciling the cognitivist drive to assign informational content to discrete internal structures with the idea that organisms are selected for what they can do given their *entire* phenotypical constitution and its relationship to the ecological niche that the animal inhabits. Information, or meaning, is best understood in these circumstances as a relationship between the behaving animal and its surroundings not as structure 'frozen' within the body.

Is the Concept of Representation Dead in the Water?

There is, to put it mildly, a lot of healthy debate within the philosophy of mind over how we might go about solving the problems of symbol grounding, representation constitution,

---

[35] This problem is also reflected in the variety of means that can be attributed to the term 'trait'. Trait can be used both in the structural sense to refer to a particular body part or organ (e.g., wings, eyes, ears), and the behavioural sense to refer to something that an organism with such and such a constitution can do (e.g., flying, night vision, echolocation). The two senses are obviously related in that, for instance, having eyes obviously has something to do with being able to see, but seeing (especially seeing in a particular way, e.g., seeing a particular sort of predator) relies on the existence of many other body structures that enable that activity (for a related discussion of sensory systems see Gibson, 1979/1986).

and content assignment. It is tempting for the critic of cognitivism to suggest that this debate is an indication of disarray and the fundamental incoherence of the concepts of mental representation and natural computation. But, as I will make clear in the coming chapters, the notion of mental representation has remained alive and well in the face of the emerging interactionist ideas. More than a few theorists have argued that modified understandings of representation are called for to help us make sense of the embodied, situated, and distributed nature of cognition. Representation continues to constitute a theoretical and methodological rallying point for many researchers. It provides a scaffold for organising empirical research design, computer and robotic modelling, and theoretical explanations in the cognitive sciences. I hope to have made it clear however that mental representation cannot simply be assumed; that it must be argued for if it is to serve as a key explanatory concept. Currently there exists no widely agreed upon understanding of what a subpersonal mental representation might be and of how we might spot one if we came across one. However, there is a widespread intuition that what cognitive scientists need from the concept of mental representation is something that can act as an internal isomorphism of the world (a model) and of potential actions (a plan). Thus the popular notion of representation seems to come with much more theoretical baggage than just the fact that it should be a physical content holder. The concept of s-representation seems to demand a modular or stage-based understanding of the operation of the mind/brain.

The Sense-Model-Plan-Act Schema and Formal Task Description

If cognitivism's representations are simulation representations then cognitivism seems committed to the view that cognizers build models of their worlds and use those models to build other s-representations that serve as plans for action. In other words people and other cognizers are not just computers implemented by wetware but special *kinds* of computers. Indeed they are not even characterisable as computers that can operate independently of an operator. Cognizers are viewed as computers that adaptively control bodies in, often dangerous, environments by transforming peripheral stimuli into useful forms of internal information that can be used to decide what to do and ultimately execute behaviour. Brooks (1991b) has referred to this view as the *sense-model-plan-act* (SMPA) *approach* to modelling cognition. Cognitive psychologists will recognise this as the central schema that underpins the usual formulation of the notion of *information processing* in cognitive psychology texts. Van Gelder and Port (1995) provide a nice example of this SMPA picture:

> According to this approach, when I return a serve in tennis, what happens is roughly as follows. Light from the approaching ball strikes my retina and my brain's visual mechanisms quickly compute what is being seen (a ball) and its direction and rate of approach. This information is fed to a planning system which holds representations of my current goals (win the game, return the serve, etc.) and other background knowledge (court conditions, weaknesses of the other player, etc.). The planning system then infers what I must do: hit

the ball deep into my opponent's backhand. This command is issued to the motor system. My arms and legs move as required. (Van Gelder & Port, 1995, p. 1)

Figure 2.1 presents a simplified example of information flow and processing in a cognitive system from the SMPA perspective. This framework is largely a simplification of Holland, Holyoak, Nisbett, & Thagard's (1986) framework for understanding inductive reasoning in all sorts of cognitive systems.

Sensing involves the transformation of peripheral stimuli, such as the intensity values of the cells in the retinal array, into an internal description of the local environment. Often researchers speak of the early cognitive processes recovering the environment from degraded sense data. Cognitive psychology texts usually depict this stage of processing as involving sequential substages of sensation, perception, and recognition and identification. The product of this series of computations is a 'meaningful picture' of the local environment. This picture is passed on to the modelling stage.

Modelling, as noted earlier in the chapter, involves embellishing the product of perception with one's stored knowledge structures. In a sense modelling is an interpretation and evaluation stage and involves discovering what the percept 'means' from the agent's perspective by providing knowledge of where things are, their utility, and so on. A complex cognitive system will likely contain a number of candidate stored representations, such as schemas and scripts, that can potentially explain the deep (i.e., non-obvious) structure of the present surroundings. These alternative possible stored representations are often thought to compete with each other based upon the degree to which they best explain the present surroundings. That is, each stored representation has an associated 'strength'. The strength of a stored representation is modified in the light of the system's experiences of the accuracy of the predictions and expectations derived from the 'view of the world' given by that stored representation. When a stored representation usefully or accurately predicts how the environment will change, its strength is increased and the strengths of other competing interpretations are decreased. If a stored representation fails to successfully predict, its strength is decreased.

Planning involves moving from simply knowing what exists in the world to deciding what to do. This involves the interplay of knowledge of what is, or may be, in the world and the system's goals. Goals are also represented, typically as desired states of the world, and are also subject to modification in the light of experience. Planning is usually thought to involve a kind of problem-solving procedure where a sequence of actions are formulated for reducing the distance between the current world state and the desired goal state. This process is sometimes thought of as involving the utilisation of an off-line emulator that conducts simulation runs on various possible plans. Ultimately one plan is selected to be passed on to the final processing stage.

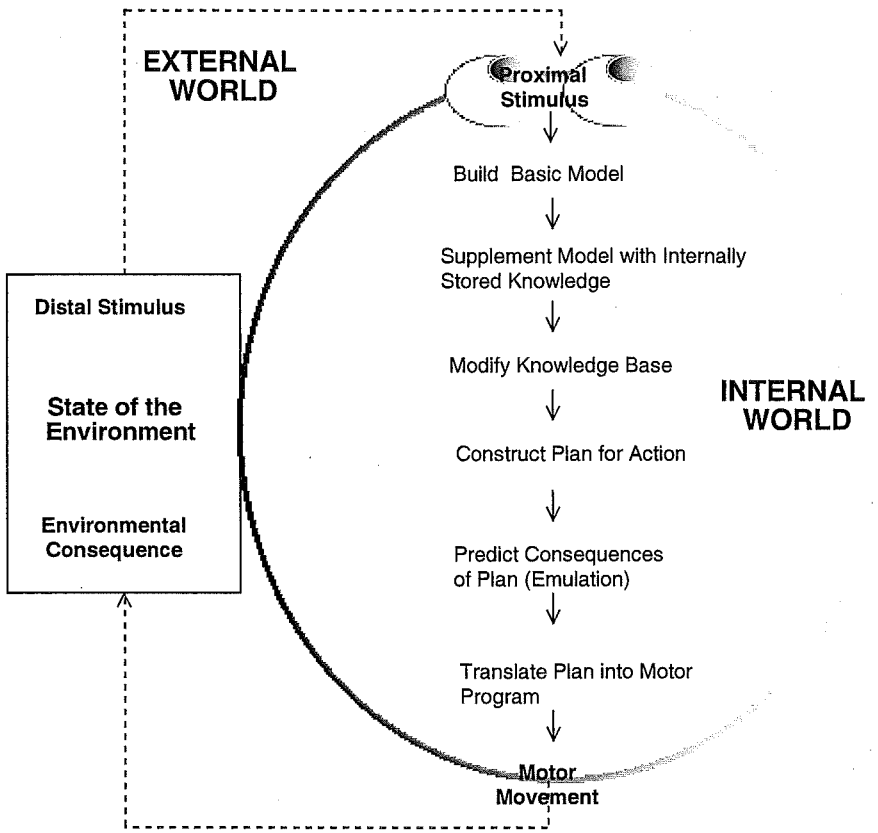

Figure 2.1 A Sketch of the Sense-Model-Plan-Act Schema

The *acting* stage consists of a translation of the plan into a motor program that instructs the various relevant body parts how to move in order to produce the appropriate behaviour.

In sum, the cognitivist vision decomposes the cognitive process into a series of 'cognitive functions'. In Haugeland's (1995) terminology each function is a *component* of the system. Within a component there exist high-bandwidth connections and interactions. Information flows thickly back and forth between the subprocesses within the component. By comparison, the interfaces between components exhibit low-bandwidth connections. They amount to information bottlenecks where the product of one stage is passed on to the next stage to be processed. Since most modern scientific endeavours require some sort of decomposition of the target phenomenon (Bechtel & Richardson, 1993) it makes sense to decompose along its natural fracture lines where the interactional bandwidths are the lowest. This, hopefully, will do the least amount of violence to the system being studied and provide the best understanding of how it works. It is hardly surprising then, that since cognitivism considers sensing, modelling, planning, and acting to be the major components of cognitive systems, the divisions of modern research follow this functional decomposition.

Deducing the Computational Blueprints using Formal Task Description

How has cognitivism arrived at the idea that the primary components of cognitive creatures are their cognitive functions? The answer seems to be that the SMPA schema seems to fall out naturally when one embraces what Hendriks-Jansen (1996) calls a *formal task description* approach to the understanding of cognitive phenomena. The formal task description approach forms a central plank of the *functionalist theory of mind* which holds that the best way to study cognitive phenomena is to understand the causal *roles* of cognitive entities. Here are two representative characterisations of functionalism:

> Functionalism maintains that mental events are classified in terms of their causal roles. Thus, a mental event would be described in terms of its role in the mental system just as a cam shaft is characterized in terms of its causal role of controlling the opening and closing of valves in a car engine. (Bechtel, 1988, p. 112)

> [T]he essence of conscious intelligence lies in the "software," in the abstract computer program, in the set of algorithmic procedures, that each normal human implements in his or her biological "hardware." ... Functionalism care[s] relatively little about exactly what processes take place inside us, so long as they implement the right input-output function. (Churchland, 1995, pp. 250-251)

Functionalists have latched onto the computer metaphor of mind in order to understand how mental states are individuated. Fundamental to functionalism is the view that psychology cannot proceed merely by examining the neurobiological structures that underpin mental activity (see e.g., Fodor, 1975). Functionalists distinguish themselves from both *type-type identity theorists* and *eliminativist materialists* (see Bechtel, 1988).

Type-type identity theorists (e.g., Armstrong, 1968; Lewis 1972, Place, 1956, Smart, 1962) propose that each *type* of mental state (e.g., all instances of seeing the colour red, all instances of the pain felt in one's leg) is identical to a particular *type* of neural state (e.g., a type-type identity theorist might argue that pain is identical to the activity in one's C-fibres). In other words, these theorists believe that there exist two equally legitimate ways of describing 'mental life': a physical neurobiological description and a higher-level psychological description in just the same way that there are two equally legitimate ways of describing the temperature of a gas: a high-level 'thermometer-based' description and a low-level description of the mean kinetic energy of the gas molecules. Such a relationship between levels is technically called a *supervenience* relation[36].

Type-type identity theory is often contrasted with *token-token identity theory* – a version of identity theory that is widely held to be more compatible with functionalism than type-type

---

[36] Sterelny (1990) notes that "[i]nformally, one level of nature supervenes on another if the supervening level somehow depends on the more fundamental level. ... A little more formally, one domain supervenes on another, if there can be no change in the first, supervening, domain without change in the second, base, domain." (pp. 43-44). The supervenience literature is large and complex and a variety of kinds of supervenience have been identified. I aim only to give a broad, informal characterisation here.

theory[37]. In philosophy the term *type* refers to a class of entity and the term *token* refers to a particular instance of a class. Token-token identity theorists hold that each particular instance of a mental state is identical with a particular instance of a brain state but that it is not possible to make the stronger type-type identity claim because our mental state concepts do not supervene on particular types of neural states. The idea that many different types of structure can serve as the substrate of a psychological state or process has come to be called the *multiple realisability thesis*. The following sort of analogy, originating in the work of Dennett, is typical of the sort of argument advanced to make this point:

> [A] clock is the *type* of thing that tells the time. OK, but notice "type of thing that tells time" is not a very materialistic way of putting things; it describes clocks in functional terms. So, let's try and define "clock" in terms of a type – mode of classification – available in physics. You will soon notice that there is no clear-cut set of physical features that defines the class of clocks. There is, in effect, no one *type* of physical thing by virtue of which all *token* clocks belong to the *type* clock. Some clocks run on springs and motors, some are digital, some have batteries, some are big, some are small, some contain sand, and so on. All clocks, of course, are physical. So token-token identity theory is true of clocks; each and every clock is a physical thing. But type-type identity theory is false of clocks; there is no physical vocabulary that can do the job our functional way of talking about clocks does. (Flanagan, 1991, p. 59)

In sum then, the majority of functionalists (those that endorse token-token identity theory) believe that the strict neural reductionist vision of mentality implied by type-type identity theory cannot serve as an adequate foundation for psychology and cognitive science.

Neurobiological eliminativists[38] argue that advanced neurobiological descriptions will make a scientific mentalistic vocabulary, at least one based upon our everyday mentalistic talk, redundant (e.g., Churchland, 1989). That is, our psychological talk of mental states such as beliefs, desires, and thoughts, even when conceptually tidied up in symbolic or information-processing terminology, will not map onto any of the real biological cognitive processes that occur in nature. Folk psychology is a theory and an incorrect theory at that[39].

---

[37] Actually the 'Australian functionalists' (see, e.g., Braddon-Mitchell & Jackson, 1996) argue that functionalism is more attuned to *type* identity theory but acknowledge that this is a minority viewpoint.

[38] Another brand of eliminativism advances the claim that a new sort of cognitive vocabulary, that is in keeping with neurobiological description and is radically different from everyday mentalistic talk, will be necessary for a scientific study of mind (e.g., Stich, 1983).

[39] One could argue that an interactionist view of cognition, like the one developed in this thesis, is also a form of eliminativism (e.g., Smithers, 1992, 1995). Keijzer (1997) sidesteps this possibility by arguing that his behavioural systems theory focuses on subpersonal issues and not the personal, mentalistic issues typical of folk psychology. Keijzer believes that we should eliminate a folk psychologically-based agent talk from our subpersonal theories but that this may be consistent with the legitimacy of a folk psychological framework at a personal level (although he provides no argument showing how this might be the case). I take a different view again and suspect, like Coulter (1979, 1983, 1989), that the 'mentalistic' view of folk psychology expounded by many philosophers of mind misreads what people are actually doing when they

By contrast with the type-type identity theorists and the eliminativists functionalists believe that some kind of informational-level of description will provide us with our best understanding of cognitive processes (see Bechtel, 1988; Braddon-Mitchell & Jackson, 1996 for reviews). They eschew neurobiological reductionism for a number of reasons including the fact that higher-level abstract description provides us with a simpler and clearer avenue for research. This approach, as Dennett (1996) notes, "promises to make life easier for the theorist by abstracting away from some of the messy particularities of performance and focusing on the work that is actually getting done." (p. 69). Perhaps the primary impetus for the functionalist perspective is the conviction that mental states and processes are defined by computational transitions and that such transitions can be implemented by a variety of physical mechanisms (the multiple realisability hypothesis). Putnam (1975), for instance, points out that it seems eminently sensible to attribute a type of mental state to agents whom we know have different underlying brain states: no two people have identical neurobiological structures, and no single person has the same neurobiological structure throughout their life, yet it makes sense to say that, in both cases, they could entertain the same mental state. For instance, it would be strange to suggest that **Believing that World War 2 ended in 1945** was only possible in a certain sort of brain, or that two different neurobiological structures attempting to entertain such a belief would, at some fine grain of analysis, actually turn out to instantiate different beliefs. And neurobiological research confirms Putnam's claim of the existence of important individual and developmental differences in the brain structures of humans. It has been shown that there are marked individual differences in the ways in which people's brains are wired up beyond a gross neuroanatomical level of description (Edelman, 1992), that slightly different regions are activated in different individuals in response to identical sensory stimulation/tasks[40] (see Elman et al., 1996, p. 298), and that people's brains undergo many changes through the lifespan. Presumably, however, all of these different neural configurations can be said to 'realise' most mental states.

---

use terms such as belief, desire, thought, memory and so on. Unfortunately there is not the space to develop this argument here; suffice to say that the alternative 'ethnopsychological approach' views 'mental state talk' as an instrument for social coordination and positioning and not an insight in to the fact that people have mental states in their heads. Note that whether the idea that people have in-the-head mental states is a sensible scientific hypothesis is another question entirely.

[40] For instance, Elman et al. (1996) discuss several pieces of research that show how human brain structure varies between individuals. MRI research shows significant individual variations in the size and shape of cortical areas even in genetically identical individuals. For instance in two genetically identical subjects the amount of cortical area occupied by the occipital cortex differed by somewhere between 3% and 7%. Another fMRI study has shown that there is a sizeable variation in the cortical regions activated by an upper and lower visual field stimulation task. There is also a long tradition of research showing individual differences in terms of handedness and hemispheric organisation for language.

Furthermore, if we were to attribute particular human-like mental states (e.g., recognising a particular researcher) to animals of different species such as chimps, dogs, aliens,[41] or artificially intelligent machines, then it becomes even more obvious that a strict structural method of individuating mental states cannot possibly provide an adequate basis for the categorisation of psychologically interesting phenomena. At the same time however it is important to realise that human brains are very much alike and indeed are very similar to the brains of other primates and mammals (Elman et al., 1996). Thus, there exists little empirical evidence to support the strong functionalist claim that the similar mental states *can* be realised by radically different physical states such as computer hardware, pieces of string, old Coke cans, and so on. That said, there is evidence for the weaker claim that we do attribute the ability to, for instance, believe that **John Lennon was a Beatle** to people with differently structured brains. And perhaps more interestingly, it seems to make sense to say that non-humans and humans (who have even less similar brain structures than two different people) can share mental states such as 'recognising a particular person'.

Refining Functionalism

An important implication of the dominance of the functionalist and token-token identity views of the mind-brain relationship is that many cognitive scientists, especially those in traditional experimental cognitive psychology, artificial intelligence, and philosophy of mind, have just *assumed* that the brain is some sort of computational-representational device in which the mind is realised without paying much attention to the neurobiological plausibility of such a view. For instance, cognitive psychologists have, until recently, used data gathered from reaction-time, discrimination, and accuracy studies to model cognition almost purely in terms of particular types of information flow and processing (e.g., Palmer & Kimchi, 1986). Several factors have led to a number of criticisms of this sort of approach and the basic computational functionalism on which it is based.

One classic objection to functionalism was made by Block (1978). He argues that different kinds of functionalism are either too liberal or too chauvinist in their assignment of mentality to systems. In *commonsense* (or *folk psychological*) *functionalism* any system that can be understood to mirror the causal relationships between commonsense psychological inputs, internal states and outputs is considered to be a cognitive system[42].

---

[41] Philosophers of mind seem to be inordinately worried about the repercussions of Martian mind/brain physiology on cognitive science. Philosophers tend to assume that Martians will have a radically different physical basis to their mind/brains.

[42] Commonsense inputs, outputs, and internal states are simply perceived entities (e.g., a person sitting down), mental states (e.g., beliefs, desires etc.), and behaviours (e.g., catching a ball) as we conceive of them in the everyday, folk psychological sense.

Block (1978) argues that systems such as the Chinese nation, or a collection of trained fleas, might mirror these relationships (i.e., duplicate the causal roles and relationships of mental systems) but that we would be unlikely to want to think of them as cognitive systems. In particular, he claims that the experiential aspects of mentality (i.e., qualia) cannot, on the face of it, be captured by such systems. By contrast to commonsense functionalism, *psychofunctionalism* describes the inputs, outputs, and internal states of a cogniser in neurobiological, physical, or technical psychological terminology. Block argued that such explanations must be couched in terms of our understanding of the workings of particular examples of cognisers (e.g., humans) and may thus be overly chauvinistic in their assignment of mentality to different systems. For instance, a psychofunctionalism based upon the kinds of internal transitions found in the human brain would withhold mentality from intelligent robots or Martians whose mental hardware worked in an unhuman manner.

Block's (1978) arguments have had a significant impact upon the ways in which philosophers have thought about the sufficiency of functionalist interpretations of mental states and their relationship to the physical substrate (see Bechtel, 1988 for a summary of some of the attempts at saving functionalism from the problems Block identifies). However, functionalist thinking has also been challenged by a number of empirical developments within cognitive science.

First, the development of a strong connectionist alternative to the traditional classical symbolic model of cognition has led a number of researchers to suggest that we *do* need to pay close attention to peculiarities of the physical substrate of cognition. Rumelhart (1989) has pointed out that arguments for the functionalist claim of the primacy of an informational/functional level of analysis that derive from an examination of computers are misleading.

> It is ... often pointed out that we can learn very little about what kind of program a particular computer may be running by looking at the electronics ... It doesn't matter whether we use an IBM or an Apple, the essential characteristics are the same. This is a very misleading analogy. It is true for computers because they are all essentially the same ... When we look at essentially different architecture, we see the architecture makes a good deal of difference. It is the architecture that determines which kinds of algorithms are most easily carried out on the machine in question ... It is thus reasonable that we should begin by asking what we know about the architecture of the brain and how it might shape the algorithms underlying biological intelligence and human mental life. (Rumelhart, 1989, p. 134)

Other connectionist-friendly theorists have claimed that connectionism has blurred the traditional boundary between implementational substrate (neurobiology) and functional description (psychology) (e.g., Churchland, 1989; 1995; Clark, 1993; 1997). Clark (1997) notes that

With the advent (or rebirth) of connectionist models, [the beliefs associated with the multiple realizability thesis] began to change. These models were deliberately specified in a way that reduced the distance between the computational story and the broad nature of neuronal implementation. ... [A]s connectionism matured ... a real synthesis of the computational and neuroscientific perspectives looked to be on the cards (p. 129).

The second important factor in the changing view of the mind-brain relationship has been the rapidly increasing interest in neurobiology and the consequent genesis of fields such as cognitive neuroscience and cognitive neuropsychology (e.g., Sarter et al., 1996). A number of researchers have argued that we need to pay much more attention to the structure of the brain before we can have an adequate understanding of the mind (see, e.g., Churchland & Sejnowski, 1992; Edelman, 1989, 1992; Koch & Davis, 1994; Reeke & Sporns, 1990). This has had the effect of reintroducing a moderate form of eliminativism into current cognitive neuroscience.

Thirdly, there has been an important introduction of an evolutionary perspective into basic cognitive science and cognitive psychology by those who claim that the basic functional analysis of psychological processes cannot be undertaken without an understanding of the selection pressures that a species has faced in its evolutionary past (Buss, 1995; Barkow, Cosmides, & Tooby, 1992; Cosmides & Tooby, 1994).

Lastly there has been a recent move toward appreciating the role played by non-neural body systems in cognitive activity (e.g. Damasio, 1994; Dennett, 1996). Dennett (1996) notes that "... it's almost standard for functionalists to over-simplify their conception of [the] task [of individuating by function rather than structure], making life *too* easy for the theorist ..." (p. 69) because they do not take into account "... [t]he fact that your nervous system ... is not an insulated, media-neutral control system - the fact that it "effects" and "transduces" at almost every juncture - forces us to think about the functions of their parts in a more complicated (and realistic) way." (p. 77)

With each of these criticisms computational functionalism has been modified (when it has not been rejected outright) in order to account for the particular concerns raised. With each modification functionalism has been come less and less intertwined with the multiple realizability thesis and more concerned with describing, in abstract terms, the actual relations that hold between neural structures, other body structures, and the environment.

Formal Task Description

The functionalist orthodoxy, even in its more refined forms, encourages the cognitive scientists to conceive of cognitive activities as neural processes that should be described at an abstract informational level of analysis. So when a cognitive scientist attempts to understand how an agent might carry out a particular cognitive activity, or a roboticist designs a machine to accomplish a certain task, he or she resorts to, what Hendriks-Jansen

(1996) calls *formal task description*[43]. Basically speaking this is the strategy of hypothesising that the mechanisms underlying cognitive activity can be discovered using a logical or algorithmic breakdown of tasks (the behavioural functions of particular activities) into a series of subtasks. Each subtask is thought to be implemented by a particular component of the agent's inner machinery. Indeed this is exactly how programmers usually write programs. So functionalism seems to encourage the cognitive theorist to think 'how could I program a computer so that it could carry out this cognitive task?'.

Van Gelder (1995) gives a nice example of this kind of thinking as it might have been applied to the problem of building a governor – a device for modulating the amount of steam needed to smoothly run an engine. Van Gelder suggests that the computationally-oriented engineer would adopt the classic cognitive science approach of breaking the overall problem task into smaller components and thereby produce an algorithm that abstracts out the 'information processing tasks' that the computational governor must do.

1. Measure the speed of the flywheel.
2. Compare the actual speed against the desired speed.
3. If there is no discrepancy, return to step 1. Otherwise,
   a. measure the current steam pressure;
   b. calculate the desired alteration in steam pressure;
   c. calculate the necessary throttle valve adjustment.
4. Make the throttle valve adjustment.
5. Return to step 1

(Van Gelder, 1995, p. 348).

Van Gelder goes on, in typical cognitivist fashion, to suggest that the computational governor that implements these steps probably possesses "a tachometer (for measuring the speed of the wheel); a device for calculating speed discrepancy; a steam pressure meter; a device for calculating the throttle valve adjustment; a throttle valve adjuster; and some kind of central executive to handle sequencing operations." (p. 348). It is interesting to note that the actual mechanical design of the governor is nothing at all like this. In chapter 5 I examine this example in much more detail. Although van Gelder's example is perhaps a little simplistic,[44] it does underline the important point that the formal task analysis

---

[43] Van Gelder (1995) refers to a similar strategy as homuncular explanation and Clark (1996, 1997) calls it componential explanation.

[44] For instance, a formal task analysis will typically be constrained by some kind of empirical observations or experiments. In cognitive science behavioural performance, as measured by, say, chronometric experiments, and neural activity pattern analysis, via brain imaging, might provide extra constraints on the hypothesized computational description. It is important to realise, however, that formal task descriptions are often constructed *prior* to such empirical work and that the empirical results often get squeezed into an

approach of functionalism seems to encourage a particular way of thinking about explaining cognition that does not provide a fail-safe strategy for constructing an adequate model of the processes involved in a particular cognitive activity.

More specifically, there are two problems that arise from adopting the formal task description method when deducing the inner workings of the cognitive computer. The first problem is that the formal task description approach does not work in cases where an activity is an emergent feature of a behavioural system. Thus exclusive reliance on formal task description effectively renders emergent functionality invisible. The second problem is that formal task description lends itself to two incompatible readings. One reading portrays the formal description as an abstraction of lower level in-the-head processes. The second reading implies that the description is a description of the tasks that the system as a whole carries out in the world. Although incompatible, these two readings are unfortunately conflated in the literature. Let's look at each of these problems in turn.

Problem 2: The Emergent Functionality Problem

The first problem with the formal task description approach is that situated roboticists and other ESD thinkers have shown that the decomposition by function approach is not the only possible way of building an agent that performs a particular kind of task. And this implies that the functional decomposition that derives from a formal task analysis is not the only way natural born cognizers (people, other animals, even Martians) might be put together. I will discuss these interactionist insights in the rest of this thesis. For the moment all we need to appreciate is that functionality can be *emergent* - indeed it can be *interactively* emergent. All this means is that the components of the agent may not correspond to sub-machines that carry out logically deduced sub-tasks. Rather, activity may arise from the operation of a collection of behavioural layers that interact with the local environmental structure (see chapters 3 and 4 for how this may occur). The variables that seem to be explicitly used by the agent (in the light of our formal task analysis) in fact only exist in the folk psychological interpretation of the observer (Smithers, 1992). More importantly, emergent functionality is not only a possibility in natural cognitive systems but, according to Hendriks-Jansen (1996), it would be a miracle if natural selection ever produced a system made up of subsystems equivalent to the subtasks in a formal task decomposition. Natural selection operates on evolutionarily successful behaviour-in-the-

---

existing computational interpretation. One could argue that this is standard scientific hypothetico-deductive practice where hypotheses are empirically tested and modified in light of the results of research. In addition to being sceptical about the adequacy of the hypothetico-deductive view of scientific method (see, e.g., Haig, 2000) I believe that much, but not all, work within mainstream cognitive science does not so much *test* the empirical adequacy of formal task descriptions as use them as unanalysed explanatory scaffolds for empirical findings.

world not on the components of an environmentally-isolated machine; and the latter would have to be the case for evolution to select a creature's innards that contained physical correlates of a task or its logically decomposed subtasks. The implications of this interactive emergence are profound for engineering (designing and building robots) and reverse engineering (attempting to understand the generation of behaviour by natural agents) in cognitive science. If a formal-task analysis cannot reliably establish how to best put together an autonomous agent, then we are left with a serious methodological problem.

Problem 3: The Incompatible Readings Problem

The emergent functionality problem suggests that formal task analysis is merely a bad explanatory heuristic for cognitive science and psychology. I want to suggest, however, that the reason this is so is because formal task analysis and, more generally, functionalism involves the conflation of two distinct understandings of cognitive activity. This is perhaps best illustrated by an analysis of Marr's (1982) popular approach to understanding cognitive systems.

Marr's Three Levels for Understanding Information-Processing

David Marr (1982) describes three different levels of analysis required when trying to understand how any information-processing machine works. Such machines include animals, animals' computational subsystems, such as their visual systems, and, if Marr is to be believed, cash registers.

Marr draws a distinction between three levels of description: a computational theory level, a representation and algorithm level, and a hardware implementation level. He suggests that each level provides a particular perspective on an information-processing problem and thus comes complete with its own set of questions. Marr (1982) summarises these in the following way:

Computational theory: What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?

Representation and algorithm: How can this computational theory be implemented? In particular, what is the representation for the input and output, and what is the algorithm for the transformation?

Hardware implementation: How can the representation and algorithm be realized physically? (based on figure 1-4, p. 25).

One common interpretation of Marr's view is to simply equate the representation and algorithm level with an abstract description of the physical hardware that constitutes the cognitive agent's mind/brain, and the computational level with an even more abstract view of the representation and algorithm level. By analogy, the representation/algorithm level would be akin to describing the running of a computer at the level of a particular piece of programming and the computational level would be akin to a higher level, logical 'flowchart' description of what the program (and ultimately the computer) does. Marr

(1982) implies this sort reading when he writes that the abstract computational theory should be understood as characterising the performance of the device as "a mapping from one kind of information to another, the abstract properties of this mapping are defined precisely, and its appropriateness and adequacy for the task at hand are demonstrated." (p. 24). We can call this *the process interpretation* of Marr because here he seems to be suggesting that each level provides a different window on the physical processes that go on within an information-processing device.

However, if we are mindful of the fact that high-level descriptions are often *functionally-inspired* there is a risk that high-level terms will not map onto low-level physical ones, especially *localised* (i.e., in-the-head), low-level terms. A 'high-level description' of a horse as a means of transport will not reduce to a physiological description of the horse because the property of 'being a means of transport' has as much to do with how a horse is *used* and *what it does* as it does with its physical structure. The same sort of problem may exist with high-level descriptions of 'mental modules' that are described as 'face recognisers' or 'explicit memory stores'.

This 'irreducibility' of levels seems implicit in another interpretation of Marr's (1982) levels of description. In this alternative *task interpretation* the computational level refers not to an abstract level of description that can potentially map onto 'lower' levels but rather to a description of the problems or tasks that an agent (or device) faces and that are 'solved' by particular subsystems within the cognitive agent. This view of the computational level is postulated in order to provide top-down constraints on our investigations of the workings of cognitive subsystems. This seems to be what Marr is getting at when he describes the computational level in terms of the kind of functions the device is built to perform. For instance, he believes that in the computational theory of vision "the underlying task is to reliably derive properties of the world from images of it." (Marr, 1982, p. 23). We need a computational theory in order to help us understand what the device being investigated does. He uses the following analogy to illustrate this point.

> [T]rying to understand perception by studying only neurons is like trying to understand bird flight by studying only feathers: it just cannot be done. In order to understand bird flight, we have to understand aerodynamics; only then do the structure of feathers and the different shapes of birds' wings make sense. (Marr, 1982, p. 27)

Evolutionary psychologists are attracted to this interpretation of the computational level because it accords with their beliefs that evolutionary analyses (theories of adaptive function) provide vital constraints on our theories of the nature and functioning of the cognitive subsystems.

> ... theories of adaptive function ... are an indispensable methodological tool, crucial to the future development of cognitive psychology. ... To isolate a functionally organized mechanism with a complex

system, you need a theory of what function that mechanism was designed to perform. (Cosmides & Tooby, 1994, p. 43)

In sum there are two interpretations of Marr's computational level: a *process* view in which it is viewed as a high-level description of the workings of parts of the brain, and a *task* view in which the computational level is understood as a description of the problems or tasks that face the cognitive agent. Simply put, *task* refers to the things an animal needs to do, and *processes* refer to the ways that it does them. It is important to note that the process aspect of describing cognitive activity is usually understood in terms of *internal* (usually neural) structures and operations. Vision, for instance, is described in terms of the ways the visual system transduces electromagnetic energy and uses it to create a representation of the world. Tasks on the other hand refer to things that the *whole animal* does, needs to do, or is designed to do *in the world.* At the task level vision is understood in terms of the sorts of environmental information that the animal requires in order to see in an adaptive manner. Processes happen inside the animal, whereas tasks describe the relations between the inside and outside. Roughly speaking this distinction maps onto the subpersonal-personal distinction.

The key point to understand here is that task and process are *not* different levels of abstraction of a single system, as, for instance, the representation/algorithm and hardware implementation levels seem to be. There exists no mapping between task and process. Rather, they portray different concerns, one to do with what needs to be done, and the other with how that will be accomplished. This is not always, or even often, appreciated, even by Marr himself[45]. The use of *level* to label the different concerns probably does not help for this term implies a stratified view of reality. But perhaps more importantly it is likely that the use of formal task analysis in the writing of computer programs that contributes to Marr's apparent conflation of the two incompatible views. If, for example, a programmer is asked to write a program to calculate prices after tax the first step is to break down the *task* for calculating correct answers into smaller subtasks (perhaps the steps that a person would follow if calculating by hand) to create an appropriate algorithm. The overall task may be to calculate prices after tax. Say the subtasks are (1) **FIND ITEM PRICE**, (2) **FIND ITEM TAX RATE**, and (3) **CALCULATE TAXED PRICE**. Each of these subtasks may be further broken down into smaller, simpler subtasks, but they are tasks nonetheless. The potential for believing that tasks are just high-level descriptions of processes occurs when the algorithmic redescription of the task is then used to produce the routines in a program that are ultimately instantiated in the physical structure of a computer. Here it appears that

---

[45] For instance, see Marr's (1982, pp. 22-24) discussion of how his three level approach can be applied to the information-processing operations of a cash register. Importantly a cash register is rather like a computer in terms of the ways tasks can be mapped onto processes.

the task description is converted into a configuration of physical structures (the process). Heil (1981) was aware of this point twenty years ago.

> [W]e must take care to avoid the error of supposing that descriptions of things done are really *indirect* descriptions of the mechanisms which get them done. This is where the use of computer models of the activities of persons seems especially pernicious. To coax a computing machine to perform a certain task, we must first say what it is we want done. This requires that we describe in a precise way the performance we have in mind. Thus, the complexity of the programmes we write may well be a function of the complexity of the descriptions with which we are obliged to supply the computing machine. If we are successful, the machine will use our description (appropriately 'coded') to carry out the task we have set for it. In this way our description of what is done does provide a characterization of sorts of the mechanism (the computing machine) which performs the task described. This, however, is simply a boring fact about the way in which we have programmed the machine. It scarcely licenses the conclusion that any device which performs the task in question necessarily does so in anything like the way the computing machine does it. (p. 327)

The concept of a subtask actually resides in an 'ontological no-man's land'. Sometimes a computer program may decompose a task in a way consonant with the series of actions a person may carry out (e.g., an algorithmic breakdown of tying shoelaces may go **PUT ON SHOE**, **GRASP BOTH ENDS OF LACE**, **CROSS OVER**, etc.). We can call this a *practical activity decomposition.* However within cognitive science the task decomposition is usually of an entirely different sort that we might call a *logical or formal decomposition.* With this form of decomposition theorists hypothesise about the logically necessary elements of a task rather than the actual temporal series of actions that a person would follow. Take the example of visual recognition; it is commonly claimed that in order to recognise an object, a representation of the object needs to be created (subtask 1), followed by some process of matching the constructed representation with those stored in long term memory (subtask 2). This is so because the logical implication of believing that 'recognition is matching' is that a system requires two entities to be compared and contrasted. This situation is unlike (most of) our experiences of visual recognition where the act of recognising an object (i.e., knowing what it is) requires little or no effort at all. That is, we do not consciously work our way through a series of steps to arrive at a solution (unlike, for instance, when we do complex multiplication using pen and paper)[46]. Hendriks-Jansen (1996) argues that formal task descriptive readings of cognitive activities arise from a basic assumption (clearly articulated by Marr, 1982) that cognitive processes are solutions to problems. In particular Marr argued that cognitive phenomena such as

---

[46] Some might argue that the reason that we feel as if there is no series of computational stages (processes) equivalent to the subtasks is because they are unconscious processes. But this need not be the case as connectionist research shows. Connectionist networks can 'recognise' an object (i.e., correctly identify a stimulus and provide an appropriate output) without the need for separate 'memory look up' and matching processes. This is a simple example of how even in-the-head emergent phenomena remain invisible to the formal task description method.

vision involve a movement from a basic kind of peripheral representation (e.g., an image) to a complex and useful deep representation (e.g., a 3D model) and that cognitive processes are the computations that transform one into the other. Hendriks-Jansen (1996) argues that:

> When Marr laid down that a computational theory should specify what needs to be computed and why before going on to specify how [in the representation/algorithm level], he did not appear to realise that, by adopting a problem-solving approach, he had already constrained his possibilities concerning what and why by making the most fundamental assumption of all concerning how. (p. 105)

The fundamental lesson we should take from this analysis is that subtasks within a formal task decomposition remain tasks: things that need to be done by the *whole* 'cognitive system' in the world. At no point do they miraculously become descriptions of the sequential internal workings of the agent: the *process* whereby the task is accomplished.

In sum it seems that the notion of explanation via formal task description raises a number of difficulties for cognitivism. There exists a disturbing lack of clarity as to what the term *function* is meant to do in psychological explanation. It is common for theorists to appeal to the notion that a neurocomputational circuit of a cognitive system has a function in the same sense that a component of an engine has a function; what might be called the *function within the system*. At first blush this seems to be what theorists like Sterelny (1990) and Bechtel (1988) have in mind when they discuss the functionalist approach to mind in the following way:

> When I talk of mental functions, the notion of function is biological; it is the same sense in which the function of the kidney is excretion and water regulation. (Sterelny, 1990, p. 11)

> [A] mental event would be described in terms of its role in the mental system just as a cam shaft is characterized in terms of its causal role of controlling the opening and closing of valves in a car engine. (Bechtel, 1988, p. 112)

Unfortunately cognitivist researchers seem to mix up this sense of function with the notion of *function for the system*. This sense of function is apparent when psychologists and their kin talk of computational subsystems having the function of solving particular problems or tasks that face the cognitive agent. Talk of neurocognitive modules that recognise faces or detect cheats confuses two distinct and complexly related senses of function, and this in turn gives the misleading impression that the cognitivist enterprise has made explanatory progress.

Conclusion: Cognitivism as a Methodological Framework

Despite the fact that I have examined some conceptual arguments regarding the plausibility and coherence of the notions of mental representation, natural computation, the SMPA schema, and the use of formal task descriptions, my main objective in this chapter has been to present cognitivism in terms of the methodological implications it has for making sense of cognitive phenomena empirically and theoretically. Cognitivism is a research

programme, or global theory, and is thus not vulnerable to direct empirical testing. What it provides us with is, not only an ontological framework for organising our observations of cognitive phenomena, but also a set of methodological constraints which specify, amongst other things, what is and is not observable, how to describe those observations, the conditions under which things are or are not observable, the instrumental means by what is measurable is measured, and the reliability of those measures (Hooker, 1987). The following is an admittedly idealised summary of some of cognitivism's key methodological constraints. I do not mean to argue that every cognitivist psychologist explicitly accepts these constraints, only that the theoretical tools and concepts of cognitivism encourage and guide the researcher toward constructing theories and designing empirical studies consistent with these constraints. The first of these guiding constraints is that researchers should think of cognition as something in the head rather than something in the world or distributed across world and agent. The notions of mental representation and the SMPA schema both derive from, and reinforce, this assumption. Christensen and Hooker (2000) make it clear how the concept of representation implicitly encourages the view that cognition is an in-the-head phenomenon:

> Of course, cognitive processes must have some important relation to the "outside", but this is finessed as representation, where the relations to the "outside" are collapsed into some kind of referring relation and the focus is on the internally characterized representational contents. The referring relation "skims over" all the detail of process organization and interactive dynamics to magically connect internal mental entities to the "outside" world. Once the general character of this reflection is specified the details are suppressed and attention focused on the "inner workings." (p. 29)

Bem and Keijzer (1996) show how the SMPA assumption reduces the role of the environment in cognition to a matter of input and output:

> What the system does in its own domain, the internal processing of the system, is what makes it cognitive, not its relations with the environment, because these relations are conceptualized merely as input and output to the system – where they come from, and what they represent, is taken for granted. (pp. 450-451)

The idea that the most important relations and mechanisms of cognition are located in the head lends itself to a downplaying of the importance of environmental structure. Therefore, the second constraining assumption of cognitivism is that the environment consists of degraded and relatively unstructured raw sense data. Given this assumption that the environment provides low grade information to the agent, theorists are naturally led to the view that purely environmentally-determined behaviour is simple and inflexible.

Since the environment is information-poor and the mind/brain is a complex device the problem of cognitive order is brought into sharp relief. In order for the mind/brain to produce adaptive behaviour the following three constraints are observed: solve the problem of cognitive order by postulating the existence of self-organising internal stimuli (mental representations); think of subpersonal components of the mind/brain as having meaningful

content; and think of the agent as a problem-solving machine that must construct deep useful representations from simple peripheral stimuli.

Finally, given this view of the relationship between the agent and the world, and of the way the mind/brain deals with this relationship, the following constraint emerges: cognition should be explained by discovering the problem an agent, or a subsystem of an agent, faces. This involves working out what information an agent starts with and what kind of information they require to produce the appropriate behaviour. One then constructs an algorithm that will get from the former to the latter. Empirical research is subsequently directed toward assessing which one of several hypothesised algorithms best accounts for the performance of participants in different kinds of experiment.

An Interactionist Alternative?

There is at least one other basic orienting picture that we can use to understand the production of adaptive and intelligent behaviour and that, I believe, underpins the frameworks of non-cognitivist frameworks such as ecological psychology and ESD interactionism. It goes something like this: Adaptive behaviour is fundamentally about possessing sensitivity to the nuances of the environment regardless of how distant those environmental features are from the agent. If there is a predator bounding toward you it pays to flee. And if there is *possibly* a predator nearby (because 'ecological laws' link particular environments with the presence of particular creatures) it also pays to keep clear (e.g., dense forested areas are 'cues' to the existence of potential danger. In an animal's *Umwelt* a place may look dangerous even though they see no predator). In other words, the alternative framework suggests that, contrary to the representational approach, it makes sense to build *situationally-determined systems* and not systems that use internal stimuli for initiating action. If this is the case then we need to view the bodily and neural resources of agents as enabling a complex kind of environmental coordination system; a system that produces appropriate behaviour in response to the environmental furniture that is about them *even if that furniture is out of immediate perceptual/causal contact* (recall that it is this lack of perceptual/causal contact which lies at the heart of the problem of cognitive order and the cognitivist claim that representations must be postulated as mechanisms that bridge superficial appearances and adaptive interpretations of the environment). The world is full of regularities that hold between perceivable things and things that are hidden or are potentially there. In order to produce *adaptive* behaviour agents need to be sensitive to the implications of these regularities. It is not uncommon for these regularities to change or be changed. Thus many, but by no means all, animals must have the capacity to update their sensitivity to changing regularities in order to survive (see Keijzer, 1997, 1998a; Christensen & Hooker, 2000, in press, for a similar approach).

No doubt one could argue that some notion of representation can be rescued from such an analysis. Indeed there have been any number of attempts in recent years to formulate more interactionist-friendly versions of mental representation and I will discuss some of them in coming chapters. I think however that there are two good reasons why we should be sceptical of claims that some modified form of mental representation is in order. First, such attempts really weaken the view of representation as an environmentally impervious content container and turn it into something that it is not - a structure that simply *enables* cognitive activity (see also Keijzer, 1997). The second reason is a pragmatic one. If representation-talk can be dispensed with in theoretical and empirical research then why bother preserving it? Indeed if representational thinking can lead us methodologically astray, as I have argued, letting it go may well free us up to pursue more fertile lines of inquiry. Both cognitivist and interactionist frameworks are consistent with the idea that animals have *knowledge* of things in their surroundings and the relations that hold between those things (that is, both are cognitive frameworks) but only cognitivism requires that the knowledge be encoded in the structures of the organism rather than simply being apparent in the animal's behavioural abilities.

For now we need to set aside the peculiar kind of behavioural coordination that 'grown humans' (i.e., linguistically and cognitively sophisticated people of, perhaps, ages five and higher) exhibit by thinking about things 'in their heads', consciously formulating plans, and imagining alternatives, hypotheticals, and counterfactuals. This may seem to be asking rather a lot, given the fact that these are precisely the sorts of activities that form the core phenomena studied by cognitive psychologists, but I hope to convince the reader by the end of this work that such abilities rest upon a basic cognitive bedrock that needs to be understood in a non-representationalist manner *first*.

Does the Concept of Representation Lead us Astray?

In this chapter I have suggested that the concept of representation represents the tip of a theoretical iceberg that encourages us to think of an agent's adaptive behaviour as enabled by a system that extracts information from the environment and encodes it so that it can be used within an inner analytic arena. It is a picture reminiscent of a general and his aides being secreted in a bunker receiving reports about activity in the field, examining them and formulating plans of action, and then sending out instructions for those on the front line to follow.

The three assumptions, the representational-computational theory, the sense-model-plan-act schema, and formal task description, constitute a mutually reinforcing set of principles for cognitivism. Cognitivists want to explain complex organisms' time-space distancing powers in terms of physically instantiated knowledge structures that reside inside those organisms. Because these organisms do not have contact with the appropriate

environmental stimuli they must possess internal contentful structures that can stand in for those 'absent' stimuli. These are subpersonal mental representations. Because complex organisms must behave adaptively most of the time, they need to be able to sense their changing environments and then supplement what they sense with their internal knowledge. This involves the building of inner models. In order to act adaptively internal plans for action also need to be constructed.

For many within psychology and cognitive science this seems like the only way that cognition can be adequately explained. Indeed, for some time it has been argued that cognitivism is 'the only game in town'. Newell and Simon (1981) sum up this attitude nicely when they write that "[t]he principal body of evidence for the symbol system hypothesis ... is negative evidence: the absence of specific competing hypotheses as to how intelligent activity might be accomplished whether by man or machine." (p.50). There have existed a number of alternative explanatory frameworks within psychology and the other social sciences throughout the current cognitivist period. But recently there seems to have been a concerted effort to construct genuine, robust, alternative explanations to those espoused by cognitivism. To be sure, cognitivism is still in its ascendancy within the cognitive sciences, and especially in psychology, but the claim that the cognitivist framework must be true because there is no plausible alternative is now, more than ever, untrue. Indeed, as I intend to show in the coming chapters, it may turn out that the interactionist approach provides a better explanation of cognitive activity than cognitivism does.

3. Embodied, Situated, and Distributed Interactionist Influences

The idea of environment is a necessity to the idea of organism, and with the conception of environment comes the impossibility of considering psychical life as an individual, isolated thing developing in a vacuum.

John Dewey (1884/1969, p. 56).

Introduction: The Emergence of ESD interactionism

In the past few years there has been an explosion of research that seeks to study cognitive activity by examining the ways in which embodied agents relate to and interact with their environments. This change of emphasis has moved the 'search for cognition' from the inside of the skull to an exploration of the relations that hold between animals (made up of bodies and brains) and their surroundings. More recently still, several theorists have tried to provide frameworks for situating the many claims and findings that derive from the work of researchers in over a dozen disciplines as well as teasing out the implications these new ways of looking at things have for established notions of cognition. These people include Clark (1997), Keijzer (1997), Hendriks-Jansen (1996), McClamrock (1995), Rowlands (1999), and Clancey (1997). By and large this revolution has yet to impact on mainstream psychology, although the gradual uptake of dynamical systems ideas throughout psychology presents a significant counterexample.

The primary focus of this *Interactionist* perspective is in the theoretical and empirical investigation of cognition as an embodied, situated, and distributed (ESD) phenomenon. Sutton (1998) summarises the aim and promise of this approach with an aquatic example:

> Dolphins and bluefin tuna ... are just not strong enough to swim as fast as they do. Their abilities derive not from great evolved strength but from a capacity to treat their medium as enabling, rather than constraining, their motions. As well as exploiting aquatic swirls and eddies to aid their manoeuvres, they actively create new and useful vortices and pressure gradients in their environment. *The world is not a jumble of obstacles against which organisms must struggle, rail, and bump, but a pool of resources with which they interact in a looping, "intricate and iterated dance".* When considering the natural, technological, linguistic, and institutional environments in which humans are embedded ... we would do well to remember that *uncanny fusions of internal and external processes, analogous to the dolphins' use of fluid dynamics, can have a startling productivity which remains invisible to investigation which stops at the boundaries of the skin.* (Sutton, 1998, p. 90, emphasis added)

Such a multi-perspective understanding of cognition takes into account behavioural, developmental, evolutionary, social/cultural, neurobiological, and ecological factors and thus draws on a number of disciplines within psychology and the cognitive sciences as well as several fields outside of the traditional 'big five or six disciplines' of cognitive science. To get a sense of the nature of this enterprise it will help to have a sense of how I use the terms *embodiment*, *situatedness*, *distribution*, and *interactionism*.

Embodiment, Situatedness, Distribution, and Interactionism

Embodiment

Unlike most cognitivist theory, the ESD approach views the living body as central to the study of cognition. This means a number of things. First, that cognition is all about moving bodies around to get things done (e.g., Keijzer, 1997). Second, this means that the role of the nervous system is, first and foremost, to 'control' and modulate bodily activity. Bodies are complex 'intelligent' systems that are coupled to nervous systems, not commanded by them (e.g., Beer, 1995a, 1995b). Third, living things, being the 'possessors' of bodies, are primarily interested in surviving – in preserving their own organisational integrity (through a process known as *autopoiesis*) and in maintaining adaptive contact with their surroundings (a process known as *structural coupling*) (e.g., Maturana & Varela, 1980, 1988; Varela et al., 1991). Fourth, since the primary purpose of cognition is 'making one's way in the real world', more advanced, reflective, and abstract forms of cognition are derived from, constrained by, and grounded in more basic forms of 'embodied cognition' (e.g., Glenberg, 1997; Lakoff, 1987; Johnson, 1987).

Situatedness and Distribution

These two concepts are often used interchangeably in the literature to refer to the idea that all, or nearly all, cognitive activity is made possible by resources that occur beyond the cognitivist mind/brain. There are subtle differences of emphasis in the use of these terms that derive from the different intellectual environments from which they have emerged (see the following section on *Cognition in the Wild*). Here I will use *situatedness* to refer to the nature of the cognitive agent and their activity. A situated approach to the study of cognition and action claims that activity is primarily the product of the moment-by-moment interactions of the agent with the furniture and inhabitants of their surroundings. The agent's primary skills and abilities include recognising patterns and using and manipulating their surroundings. Higher cognitive abilities, such as 'off-line' reflection and effortful planning are grounded in these basic perception-action abilities. This stands in contrast to the cognitivist view of the individual as primarily a symbol-manipulator that models its surrounds and plans what to do before 'executing' behaviour. By contrast, when I say that cognition and action are *distributed,* it is in the sense that a full explanation of many activities requires us to trace 'flows of information processing' beyond the skull and into environments of tools, practices, and other agents. Distributed cognition analyses focus on the ways in which external resources are used to tackle different tasks and the ways in which environments are constructed so as to support certain kinds of cognition and action. Indeed, it is often claimed that the environment contains entities that play important roles as vehicles for thought. To focus only on what lies under the skin or the skull is to do

violence to a natural and integrated functional system. Thus distributed cognitive analyses focus on the agent-environment system as the primary level of analysis.

Interactionism

Since the primary unit of analysis for the ESD theorist is the agent-environment system, there exists a great deal of interest in understanding the ways in which bodies, brains, artifacts, and environments interact with each other. Cognitive activity emerges from the interplay of (often) simple reflexes and the layout of the local environment and, importantly, does not exist preformed in either the agent or the world (Hendriks-Jansen, 1996). The interactionism advocated here is not the *naïve interactionism* of much developmental science critiqued by the likes of Pinker (1997, pp. 33-34) and Oyama (1989) but what Oyama calls a *constructive interactionism.* The contrast is between a view of traits or behaviours being an outcome of the 'mixing' of environment and innate structure and a position that recognises the 'informational' and constructive roles of multiple resources.

With these terms roughly characterised we can now turn to the basic structure of the chapter. In this chapter I aim to discuss and summarise the ideas and findings from several broadly cognitive scientific fields of research. The intention is to provide the flavour of the common aims and claims of this rather disparate collection of research endeavours. The next chapter is an attempt at distilling out the major shared themes of the research programmes in this loose coalition. The aim is to make explicit the basic explanatory commitments of an Interactionist perspective. This will lay the foundation for the explanatory framework I begin to develop in chapter 5.

Influential Fields

Progenitors of the ESD approach include, in one way or another, John Dewey (see Clancey, 1997), William James (see Good & Still, 1998), Lev Vygotsky, Jerome Bruner, Eleanor and James Gibson, Wilhelm Wundt (Cole & Engestrom, 1993), a number of 'Continental theorists' including Merleau-Ponty (Varela et al., 1991), Heidegger and Husserl (Wheeler, 1996; Kadar & Effken, 1994; Good & Still, 1998), the Gestalt psychologists (e.g., Lewin and Koffka see Valsiner, 1997; Good & Still, 1998), and the ethologists Von Uexküll, Lorenz, and Tinbergen (Hendriks-Jansen, 1994/1996, 1996).

Modern inspiration for ESD interactionism derives from a number of quarters, including the work of ecologically-oriented researchers (e.g., Costall & Leudar, 1996; Gibson, 1979/1986; Goldfield, 1995; Ingold, 1993), situated action and situated cognition researchers (e.g., Clancey, 1997; Suchman, 1987), distributed cognition researchers (e.g., Hutchins, 1995a, 1995b; Salomon, 1993), the growing situated and embedded movement in artificial life and 'the new artificial intelligence' (e.g., Brooks, 1991; Steels, 1995),

externalist philosophy of mind (e.g., Clark & Chalmers, 1995; McClamrock, 1995), developmental systems theory (e.g., Oyama, 1985), Vygotskian-inspired developmental psychology (e.g., Berk & Diaz, 1992; Lock, 1980; Nelson, 1996; Valsiner, 1997; Vygotsky, 1930-1935/1978, 1934/1986), radical approaches to connectionism (e.g., Bechtel, 1997; Clark, 1993; Elman et al., 1996; Globus, 1992; Pfeifer & Verschure, 1992a, 1992b), autopoietic and autonomy analyses (e.g., Christensen & Hooker, 2000, in press; Maturana & Varela, 1980, 1988), dynamical theories of cognition and development (e.g., Port & Van Gelder, 1995; Thelen & Smith, 1994), as well as psychological, anthropological, and archaeological studies of the evolution of language and modern human behaviour (e.g., Deacon, 1997; Lock, 1999; Noble & Davidson, 1996). Although all of these fields have made a mark on the nascent interactionist approach, several fields have been of particular importance and form the basis of my treatment in this section. The domains that perhaps best illustrate the fresh perspectives and theoretical innovations of the interactionist approach are: situated robotics; interactive vision; dynamical systems approaches to development and cognition; studies of socially distributed and situated cognition; and socio-historical studies of child development.

Situated Robotics

Situated robotics (a.k.a. autonomous agents research, the animat approach) forms a central component of, what has been called, the *new AI* as well as one of the main divisions of the rapidly growing field of *artificial life* (Alife) (see Boden, 1996; Langton, 1995, 1992/1996). At first blush it may seem odd that robotics, being a field deeply anchored in computer technology, could supply us with what Hendriks-Jansen (1996) calls 'existence proofs' that bring into question some of the fundamental assumptions of the representational-computational framework. However researchers such as Rodney Brooks and his MIT colleagues and students (Brooks, 1991a, 1991b, 1997; Brooks, et al., 1998; Brooks & Stein, 1994; Mataric, 1991), Steels (1995, 1998), Beer (1995a, 1995b), Pfeifer, Verschure, and Scheier (Pfeifer & Verschure, 1992a, 1992b, Pfeifer & Scheier 1999), Smithers (1992, 1995), and Husbands, Harvey, and Cliff (Harvey, Husbands, & Cliff, 1994; Husbands, Harvey, & Cliff, 1995) have revolutionised the way in which robot engineers have gone about designing and building 'artificial agents'. This revolution has been inspired primarily by the many sticky problems that robotics researchers have confronted when they have tried to construct robots in keeping with the cognitivist framework. It is significant that the solutions and new ideas brought to bear on such practical problems have provided a basis for many non-roboticists investigating the implementational and architectural foundation for a noncognitivist psychology (see, e.g., Clancey, 1997; Hendriks-Jansen, 1996; Keijzer, 1997).

Problems with Traditional AI

Traditionally robotics design has taken the form of a kind of attempted implementation of cognitivist theorising. Following the lead of theorists such as Marr (1982) and Newell and Simon (Newell, 1990; Vera & Simon, 1993) roboticists have concentrated on trying to recover the world inside the robot in a rich representational format that contains explicit and meaningful (to the robot) decomposition of its surroundings into objects and places and (occasionally) events. This re-presented internal model of the surroundings is then analysed, plans for action (movement) are formulated and finally executed. The similarities of this design procedure and the cognitivist model of mind (see chap. 2) should be obvious. Roboticists have adopted the *decomposition by function* approach for building their robots that derives from the sense-model-plan-act schema (Brooks, 1991b). Right from the start engineers have struck problems attempting to implement this kind of design strategy. These problems have included:

The Frame Problem(s)

There are numerous formulations of the *frame problem* (see the various papers in Pylyshyn, 1987) but the basic idea has been nicely summarised by Janlert (1987) as the problem of "finding a representational form permitting a changing and complex world to be efficiently and adequately represented." (pp. 7-8). The trouble with a classical AI representational system is that *everything* it does must be based upon something that is explicitly represented in its cognitive innards. Thus, if it is to accurately and usefully act, it must have some method of maintaining a continuous and meaningful mapping of the world so that it can work out what to do (i.e., plan) to achieve whatever goals it has. A representational system will need both a large number of representations (a huge stock of 'common sense') in order to know what to do in many types of circumstances and a method of deciding which of its many knowledge representations are relevant to the current situation. More specifically, a representational robot will need to assess the consequences of its actions in order to plan how to act. The central difficulty is in constructing some sort of procedure for working out which of the many consequences are actually relevant to that planning. A robot that systematically decides that, if it moves behind a wall to avoid some danger, then the sun will keep shining, gravity will not change, the wall will not turn into a gas, and so on and so forth, will be preoccupied with all sorts of irrelevancies and likely yield to the danger before anything useful can be concluded. One could argue that having a fast processor and a large memory may overcome the problem but this does not make the procedure any more biologically plausible. In fact it may turn out that there is just no way a classical AI system can meet this problem, fast processors and clever algorithms notwithstanding.

The Problem of Brittleness

Robots designed following the decomposition-by-function approach have also tended to be extremely brittle. That is, they have tended to be thrown off by small malfunctions or slightly noisy environments. This is especially the case as the internal complexity of the robot increases (see Keijzer, 1997, p. 117). Representational automata usually require very accurate signals to work. Encounters with novel and complex situations seem to often play havoc with their modelling and planning functions. Connectionist networks have been proclaimed as solutions to problems of brittleness but they have their own problems (see Pfeifer & Verschure, 1992a). Not the least of these is the common charge of their lack of *compositionality* (or ability to manipulate representational atoms by way an explicit syntax) (Fodor & Pylyshyn, 1988; see chapters 7 and 8). It is not at all clear that a complex representational robot, of the sort built by classical AI engineers, can be built without a compositional representational architecture.

The Instruction Problem

Keijzer (1997, pp. 117-122) points out an unacknowledged problem that engineers must contend with when figuring out how to turn internally modelled plans into actions. The *instruction problem,* he argues, is the efferent analogue of the frame problem. It involves the question of how to build a system that can execute plans in a way that is sensitive to the unique peculiarities of any particular environment. Keijzer calls the actual subtle, environmentally-specific, fine-grained activity of agents their *proximal behaviour* and the high-level description of similar examples of this *distal behaviour.* He claims that cognitivism has avoided the *instruction problem* by focusing attention on distal behaviour which is less complex looking than the actual behavioural activity of agents. Of course the proof is in the building of robots and here it does not look good for cognitivist inspired designs. A robot, for instance, may have the plan 'walk slowly forward ten paces' (its distal behaviour) but have no way of modulating such an instruction that takes into account the numerous environmental features that may affect it – things such as the unevenness or mushiness of the ground, wind speed, restraining effects of things caught on the robot's legs and so on (its proximal behaviour). In sum, Keijzer argues that it is highly unlikely that any representationally-based instruction system can ever specify the exact motor outputs that are adequate for producing the kind of behaviour required to achieve a task in varying conditions.

Attempted Solutions

In order to make progress in the face of these problems traditional AI workers have pursued a number of simplifying strategies in order to deal with the messy complexity of real-world cognitive activity. These include focusing efforts on the 'model' and, to a lesser extent, 'plan' functions of cognitive activity, leaving perception (recognition, sensing) and

action (movement, motor programs) for later (or preferably someone else). The assumption here, of course, is that a 'sensing function' will deliver something like a symbolic picture of the current environment to the 'modelling function' and that the 'planning function' can easily transduce its symbolic states into an adaptive movement pattern. Neither of these assumptions have stood the test of time. The transduction of energetic signals into 'internal symbols' and the subsequent transduction of symbols into action constitute extremely difficult, if not impossible, problems for traditional AI (see, e.g., Keijzer, 1997, chap. 3).

Another widely used simplifying strategy is that of modelling single domains of cognitive activity (as, for example, expert systems do) instead of "the full gamut of human intelligence." (Brooks, 1991a, p. 140). This strategy assumes that one day all of these independent inquiries will somehow meld together into a coherent whole[47] and that, by implication, there is no principled problem with the idea that different cognitive abilities can be studied and simulated in relative independence from each other.

A third technique is to simplify the environment that artificial agents must live in by creating toy worlds full of a relatively small subset of simplified and easily distinguishable objects and places. For instance, the world of *Shakey* (see, Dennett, 1991), the first mobile robot, consisted of a collection of carefully lit, uniformly coloured, linked rooms containing a few large, coloured blocks and wedges. The boundaries of floors and walls were marked with dark rubber baseboards. All of these feature were designed to make it *possible* for Shakey's SMPA-based modelling and planning control system to operate (see Brooks, 1991b). (Despite such assistance, Shakey could still only move at about two metres per hour). Ironically, as I will discuss in the next chapter, there is actually good reason to believe that the redundancy and complexity of the real world may actually be of *benefit* to AI engineers.

Of course the use of these simplifying strategies does not, in itself, challenge the fundamental assumptions of cognitivism. Defenders of the dominant approach can argue that a 'human-size physical symbol system' will need to be implemented in a very large, complicated, and powerful computer such as the brain (whereas the resources that AI workers have to work with pale in comparison). They can also suggest that the brain may utilise some kind of as yet undreamed of set of algorithms that can deal with the problems of traditional AI. No one can offer a knockdown rejoinder to these arguments. At the same time, however, such responses give no good reason for following a cognitivist approach to

---

[47] Actually, the standard attitude seems to be that AI should focus on these more tractable expert system issues than on building a fully rounded artificial intelligence. In a sense this is a retreat from the heady goal of building intelligent agents to one of supplementing human abilities with those of machines (see, e.g., Dreyfus & Dreyfus, 1986).

the exclusion of all others. As a growing number of robotics and New AI researchers are discovering, this is not the only way to think about the workings of agents. Indeed it seems that is not the best way either.

A New Way of Doing Things

Rodney Brooks is perhaps the most well known advocate of the situated robotics alternative to the traditional AI approach. He was one of the first researchers to outline how it may be possible to avoid many of the problems noted above. Brooks' goal has been to design and build robots that can survive unaided in a real world environment. All of the simplifications tried by traditional AI workers are rejected. His requirements for the creation of one of his 'Creatures' include the ability to "cope appropriately and in a timely fashion with changes in its environment"; that it "should be robust with respect to its environment; minor changes in the properties of the world should not lead to total collapse of the Creature's behavior; rather one should expect only a gradual change in capabilities of the Creature as the environment changes more and more ..."; it "should be able to maintain multiple goals and, depending on the circumstances it finds itself in, change which particular goals it is actively pursuing; thus it can both adapt to surroundings and capitalize on fortuitous circumstances ..."; and it "should do something in the world; it should have some purpose in being." (Brooks, 1991a, p. 145).

The 'Fast, Cheap, and Out of Control' Design Philosophy

In order to achieve these goals Brooks and his colleagues have effectively thrown out the traditional symbol system approach to designing AI machines. Gone is the emphasis on high-level cognitive skills such as reasoning and logic. Brooks (1991a, pp. 140-141) argues that organic agents have been billions of years in the making and that for most of that time the focus has been on the evolution of basic, survival-oriented, perception-action abilities, and that AI engineers should take this leaf out of Mother nature's book. His first message is to start by building simple creatures that can survive in the hurly burly of real world environments. Brooks (1991a) sees a number of dangers in trying out early prototypes in simplified environments (including simulated computer environments that simulated robots live in) consisting of features such as "matte painted walls, rectangular vertices everywhere, colored blocks as the only obstacles" (p. 150). In particular in "a simplified world ... it is very easy to accidentally build a submodule of the system which happens to rely on some of those simplified properties. ... When it comes to move to the unsimplified world, we gradually and painfully realize that every piece of the system must be rebuilt." (p. 150). It is only when a simple kind of ongoing, adaptive 'intelligence' is engineered in a real world environmemnt that greater behavioural complexity can be tackled. This is achieved by adding additional behavioural layers atop a previously successful and totally

autonomous first layer. Within this set of demands important design features have arisen within Brooks' research programme.

The Brooksian approach rejects classical AI's strategy of decomposition by function which depends on formal task descriptions of cognitive activities. Instead he advocates a strategy of *decomposition by activity* and the construction of, what he calls, *subsumption architectures.*

Robot design begins by building a basic behaviour layer. This layer will make the robot autonomous in a real-world environment, enabling it to respond in an appropriate and timely fashion to environmental events and contingencies. Typically the behaviour of this basic layer is something simple like object avoidance. The layer is made up of a collection of simple 'reflexes' realised by augmented finite state machines (AFSMs). These are tiny computers that receive small packets of 'information' from a variety of crude sensors (e.g., the position of a leg motor or a the readings of several noisy sonar returns) and manipulate it a bit before sending signals to other AFSMs[48].

Construction then proceeds in a vaguely evolutionary fashion. A new layer is attached to the basic layer, to enable the robot to engage in a new kind of activity. The components of the new layer are side-tapped into existing wires in he basic layer. These low bandwidth connections can either inhibit the output or suppress the input signals leading to AFSMs in the other layer. The layers thus cooperate and compete with each other so that adaptive behaviour is achieved. Usually making such a system work involves a lot of real-world debugging, sometimes by modifying sensors, actuators, and body parts, rather than the control system itself[49]. (Smithers [1992] argues that it is often the case that altering the bodily morphology is *easier* than trying to rectify a problem by altering the internal control system). Others layers are gradually added in a similar incremental fashion leading to a rather odd-looking layered architecture of semi-independent competing/cooperating behaviour producers. Each layer works independently, receiving its own input and making its own 'decisions' about what to do. Because earlier layers are not in any way dependent on later layers for their basic operation, Brooks-style mobots are much more robust (than their classical cousins) when confronted with physical malfunctions or difficult environments that may 'confuse' higher layers. No centralised or explicit model of the

---

[48] An AFSM is made up of a finite state machine (FSM) and a collection of other components such as a timer and registers (simple 'message storers'). A FSM is essentially a tiny computer with a limited number of possible internal states.

[49] Other roboticists have attempted to automate this debugging system by artificially evolving the robotic architectures using genetic algorithms (e.g., Harvey, Husbands, & Cliff, 1994; Husbands, Harvey, & Cliff, 1995).

world exists inside the robot, just a collection of coordinated testers, detectors, and measurers coupled to an equally large collection of motor control systems.

Brooks (1991a, pp. 152-154) gives an example of this design strategy with a simple mobot made up of three layers.

Allen

Allen (named after Allen Newell, a well-known proponent of the traditional SMPA approach) was one of the first mobots that Brooks built using a subsumption architecture (Brooks, 1986 cited in Brooks, 1991b). Brooks' aim was to build a simple robot that avoided people and other moving obstacles while also avoiding collisions with static objects in a real-world university office environment. When not avoiding things Allen could also set out to get to certain places that it had 'decided' were worthy destinations. Such a simple sounding robot has been embarrassingly difficult to build using traditional techniques. Yet Allen successfully lived in the MIT office environment using a very simple internal architecture.

The first layer (AVOID) enabled the mobot to avoid hitting static and moving objects (via a ring of twelve ultrasonic sonars). Its default activity was simply to sit still unless it detected a nearby obstacle (e.g., a wall a certain distance ahead or to the side or a person moving within a certain range). Upon sensing an object the mobot reoriented itself in order to move in an unobstructed or least obstructed direction.

With the second layer (WANDER) added, the mobot wandered about in random directions when not in 'object avoidance mode' (although it continually attempted to move as best it could in the direction that it has 'decided' to wander). Thus the WANDER layer was overridden by the AVOID layer in obstructed surroundings, and the AVOID layer was subtly modified by the 'decisions' of the WANDER layer.

The third layer (EXPLORE) enabled the mobot to look for distant places to move toward. This layer suppressed the WANDER layer when a suitable 'distant place' was located but it was still susceptible to course changes when the AVOID layer was activated by, for example, a person walking in front of the mobot. This simple mobot performed in the real world (of the MIT labs) in a robust, adaptive, and seemingly intelligent manner even though it possessed no central executive 'command and control' system that created models of the world or plans for activity.

Some More Examples

Brooks' student Jon Connell used Allen's basic architecture as a foundation for building a more complex robot called *Herbert* (after Allen Newell's colleague Herbert Simon). A further three layers were added to Allen's basic design to transform the mobot into a Coke can collecting machine. Herbert wandered the MIT labs looking for empty soft drink cans,

of which there were many (although no one mentions how Herbert could tell an empty can from a full one!). It used a laser combined with a video camera to detect cans and an arm to collect them in cluttered office environments. Herbert is described as being moderately successful (Brooks & Connell, 1986 cited in Brooks, 1991a; Connell, 1989 cited in Brooks, 1991b; see also Clark, 1998, pp. 507-509 for a description).

Brooks and his students have also built a collection of robotic autonomous walking vehicles (called *Genghis*, *Attila*, *Hannibal*, and *Boadicea*) that can traverse uneven surfaces and even for get themselves out of trouble when they tip over. The impetus for these robotic designs derives from a demand for machines that can explore distant places where remote control is difficult (due to radio transmission lags) or impossible (due to an inability to maintain radio contact) (e.g., Mars, volcano caldera). The behaviour layers in these robots consist primarily of systems for producing adaptive movement. For instance, Genghis consists of a collection of layers that implement the ability to stand up, enable the robot to walk with feedback, adjust motor activity for rough terrain and obstacles using feedback, and modulate for this using pitch and roll inclinometers (Brooks, 1989 cited in Brooks, 1991b).

Cognitive Robotics

In recent times Brooks and his colleagues (Brooks, 1997; Brooks & Stein, 1994; Brooks et al., 1998) have begun building a much more sophisticated humanoid robot called *Cog*. The goal of the Cog project is to build a human-like robot using the subsumption architecture methodology. So far Cog only has a head (with eyes and ears), arms, hands, and a swiveling torso. The main objective of the Cog research to date has been to approximate the sensory and motor dynamics of a human body by simulating human visual, auditory, tactile, and proprioceptive sensory-motor activities. However, the ultimate hope is that robots like Cog will be capable of something like human intelligence. Brooks has referred to this new research project as *cognitive robotics*. According to Brooks (1997) the key differences between cognitive robotics and his earlier behaviour-based robotics are: an increased concern with the importance of bodily form, motivation and processes necessary for 'deciding' which activity is most appropriate in a particular situation; designing systems for ensuring that the large number of behavioural layers that must exist within such systems can exhibit coherent behaviour; and implementing the ability to learn and modify behaviour in response to physical and social interactions (see also Christensen & Hooker, 2000 for a critical discussion of some of these issues). Importantly the goal of Brooks' team is to simulate human-like intelligence using the same basic non-cognitivist design philosophy that has been successfully applied to their earlier robots. This motivation is supplemented by a belief that human intelligence is inherently developmental, social, embodied, situated, and multimodal (see Brooks et al., 1998).

In sum, situated robotics research aims to design and build usually simple, mobile robots that act in an appropriate, timely, robust, and flexible manner in real-world environments. To this end situated roboticists have parted with their traditional symbolic AI kin by building their 'creatures' out of relatively independent, behaviour-producing layers rather than by adopting a centrally-controlled, sense-model-plan-act architecture. Moreover, these mobots do not store or construct models of their environments in order to plan how to act - moboticists eschew the use of mental representations (at least in the traditional sense). Rather, behaviour arises from the coupling of the mobots' simple reflex subsystems and the unique structure of the local environment. Such activity is said to be *interactively emergent* (see chapter 4). That is, the often complex-looking, seemingly goal-oriented behaviour is not prespecified (encoded, programmed, or discoverable) in either the mobot's innards or the world in which it acts. Rather, it is the outcome of a dynamic coupling of body structures and the layout of the world. The implications of this research framework are profound: of primary importance is the notion that the agent (in this case, the mobot) does not need to explicitly re-present either the outside world or its 'goals' or 'choices' in order to produce complex activity.

Interactive Vision

The field of *interactive vision* (also known as animate vision, active vision, and perceptual activity theory) has also arisen from robotics-related research (e.g., Ballard, 1991), although a number of psychologists and neuroscientists have also begun to investigate vision and other senses from an interactive perspective (e.g., Ballard, Hayhoe, & Pelz, 1995; Ballard, Hayhoe, Pook, & Rajesh, 1997; Churchland, Ramachandran, & Sejnowski, 1994; Rensink, O'Regan, & Clark, 1997; Zelinsky, Rao, Hayhoe, & Ballard, 1997). Within robotics, interactive vision researchers have found that artificial visual systems often achieve their goals better and in a more efficient manner when detectors can move about and rapidly interrogate the surroundings for needed information. Thus, artificial interactive vision systems "have anthropomorphic features such as binocularity, foveas, and most importantly, high speed gaze control." (Ballard, 1991, p. 58). From a psychological perspective the interactive perspective involves the claim that our sensory mechanisms are not designed to construct detailed representations of our surroundings (as in, e.g., Marr, 1982), but rather are viewed as being made up of anatomically overlapping *perceptual instruments* (*smart perceptual mechanisms* [Runeson, 1977]). These instruments are deployed to rapidly and actively interrogate the local environment for specific task-relevant qualities. Table 3.1 provides a summary of the key differences between the interactive approach and the traditional approach to understanding vision within cognitive science.

Computational simulations using the interactive vision paradigm have revealed that tasks that make massive demands on the speed and capacity of the classically designed systems can often be avoided by using computationally cheap interactive systems (Ballard, 1991).

Ballard et al. (1995, 1997) describe a simple simulation carried out by Whitehead and Ballard (1990) that illustrates this point. They simulated a simple 'hand and eye robot' whose job was to retrieve a green block from a variety of situations where the block could be found amongst, and often under, other coloured blocks. A classically designed system would typically build up a complex internal model of the situation and locate each of the blocks within some kind of spatial coordinate system. A plan would then be formulated for moving blocks off the target to other parts of the modelled environment. Whitehead and Ballard's interactive system, however, could achieve the task by applying a *deictic strategy* consisting of a repeating sequence of simple sensing and manipulation steps that required no detailed model or plan. The system used fixations on simple properties of the environment (e.g., fixate on the green block) and manipulations directed to where the system was fixating (e.g., pickup what I am looking at).

Table 3.1: Comparison of Interactive Vision and Pure Vision Approaches
(based on information in Churchland et al., 1994)

	Pure Vision	**Interactive Vision**
Task of visual system	To create a fully elaborated model of the world in the brain.	To guide action and motor control in the service of basic survival skills. *Influenced by activity in motor and other sensory systems as well as overall goals and functions.*
Scene analysis process	Whole scene is mapped out in a task-neutral fashion before analysis directed toward 'pulling out' objects of interest.	We analyse only partially elaborated aspects of the visual scene. Visual system uses saccades and small receptive field of fovea to obtain snippets of information about important aspects of local environment. Unattended objects are only minimally attended and often not visually experienced at all.
Nature of vision	Vision is largely passive. Information enters process as 'snapshot'.	Vision is exploratory and predictive. Prediction is improved by use of other modalities. Ongoing active exploration used to identify objects.
Relation to motor activity	Scene is fully elaborated before motor activity	'Motor assembling' (for eye and head movements, postural control) begins on basis of preliminary and minimal visual analysis.
Relations between stages of visual processing	Processing is hierarchical and unidirectional passing through early, middle, and late stages.	Processing is richly recurrent with action planned and feeding back into all stages of processing.
Relation to memory	Vision is informationally encapsulated; not affected by other 'cognitive functions'.	Visual learning modifies what is seen and the exploratory strategies used by visual system.
Pragmatics of research	Vision can be studied independently of other 'cognitive functions'.	Need to take into account relationships between visual and non-visual functions.

Importantly there is evidence that these ideas are not just useful for making more efficient kinds of robotic devices but also that it is very likely that humans and other animals engage in just the kinds of activity that the interactive vision approach proposes. For instance, Ballard et al. (1995) and Ballard et al. (1997) have conducted research that shows that when people engage in simple sensorimotor tasks they do not exhibit activity that is consistent with the building of detailed inner models of their environment. Rather they seem to continually manipulate it and mine it for information.

In a series of experiments Ballard et al. (1995) asked participants to manually replicate a model pattern of coloured blocks in a workspace using blocks from a resource space. The participant's eye and hand movements were tracked. Ballard et al. suggest that from a pure vision perspective we might expect each participant to first fixate on the physical model and use the information gathered to build a relatively detailed cognitive representation of the block pattern. They claim that it is well within the capacity of visual memory to remember (i.e., to store in one's head) the characteristics of eight blocks by looking at the pattern four times (i.e., holding a single two block subpattern in one's head at a time). One would expect a participant to look at the model maybe four times and look at the workspace and the resource area the rest of the time. What they found, however, was that participants fixated on parts of the model display a lot, sometimes up to eighteen times. There seemed to be a lot of 'extra' looking going on.

By far the most common pattern of fixations was, what they called, the MPMD pattern (where M stands for model, P for pickup, and D for drop). The MPMD pattern consisted of the following strategy: glance at the model (for a coloured block), glance at the resource to pickup an appropriate coloured block, glance back at the model to ascertain the block's position, and then glance at the workspace to see where to position it. They suggest that each fixation gathers just one piece of information, for example, a single block's colour or location (but hardly ever both).

Ballard et al. hypothesised that other strategies would occur when one of the features (colour or location) was known (remembered, or possibly, 'obvious'). As the task wears on there exist more constraints on possible block choices. Participants do not need to check for colour for the last block if the correct number of blocks are present in the resource area. They also suggest that repeated visits may lead to *some* extra information being picked up for longer than one cycle of block copying. They hypothesised that different (non MPMD) strategies are consistent with knowing different amounts of information about a block:

If colour is known a PMD sequence should occur – one knows the colour so there is no need to check the model for colour first. If location is known (something one would expect would be harder to remember – therefore one would expect a lower number of occurrences of this strategy) an MPD sequence should occur. One checks the model for colour, grabs

the block from the resource area, and then drops it. And if both colour and location are known a PD sequence – grab and drop – is likely.

Ballard et al. anticipated the objection that the extra eye movements may have nothing to do with the incremental, sequential acquisition of information. Perhaps people do build an internal mental model but, because eye movement is faster than hand movement, they have time to look about and, perhaps, do some checking to see that things are going okay. To test this possibility Ballard et al. ran the experiment using a collection of uniformly coloured blocks. If the objection is correct there should be the same number of 'extra' fixations in this condition as in the original experiment.

Contrary to the claims of this 'fast vision' objection they found that the number of fixations dropped dramatically in the uniform colour condition. In such a condition there is no need to ascertain the colour of individual blocks. If the one fixation-one information idea is right, the number of uses of model-first strategies should drop (MPMD, MPD) and the number of pickup-first (PMD, PD) strategies should increase relative to the first experiment. That is just what Ballard et al. (1995) found.

In another experiment Ballard et al. (1997) tried to further ascertain the nature of the information picked-up by the participants by changing the colour of one of the uncopied blocks in the model as the participant was saccading from a previous position (in the resource area or workspace) to the model. They predicted that if the participant had just detected colour information they would show some sign of noticing something was wrong. However, if the participant had just detected location information – and thus discarded any knowledge of colour (because we are assuming that participants only pick up one bit of information at a time) – they would not notice any change.

Ballard et al. changed a model block 25% of the time. The 'colour information in mind' group were those participants that had just picked up a block from the resource area and were saccading to the model area to assess location (the After Pickup condition). The 'no colour information in mind' group were those participants who had just dropped a block into position in the workspace and were saccading to the model to select the colour for a new block to pickup (the Before Pickup condition). They measured the length of fixations at the model when the block was changed/ not changed.

What they found was that participants hardly ever noticed the change in colour of the block in the model. More interestingly however, those in the 'colour information in mind' group exhibited significantly longer fixations at the model when there had been a block change than when there had not. The 'no colour information in mind' group showed no significant change. These results are consistent with the idea that colour is detected using fixations before location.

In sum, it appears to be the case that at least some visual activities involve the rapid and task-focused eye, head, and hand movement aimed at picking up very minimal pieces of environmental information. Vision, in this view, consists of many relatively rapid cycles of information collection and environmental exploration. No overall picture of the environment is ever built-up. Rather animals' sensory systems are understood as being primarily designed to detect either those bits of information that are useful for directly modulating behaviour, or those bits of information that might help find other information useful for modulating behaviour. This stands in contrast to the conventional cognitivist view that vision is primarily concerned with building sophisticated 3D models of the perceived world.

Dynamical Systems Approaches

Dynamical systems research in the study of cognition, action, and development addresses many of the same themes that have emerged in the field of situated robotics. In last decade or so a number of psychologists and other cognitive scientists have made use of ideas and tools from *dynamical systems theory* (DST) that have been used to understand patterns and processes of phenomena studied within the natural sciences (e.g., Goldfield, 1995; Kelso, 1995; Kugler & Turvey, 1987; Port & van Gelder, 1995; Saltzman, 1995; Thelen, 1994; Thelen & Smith, 1994; van Gelder, 1995, 1997). Dynamical systems theory provides rigorous mathematical and conceptual tools for understanding, describing, and explaining complex (often nonlinear and chaotic) behaviours, forms, and patterns. Although many of the psychological applications of dynamical concepts have been more metaphorical than quantitative in nature, they have resulted in some rather revolutionary ideas about the ways in which behaviour, including complex human behaviour, may be generated. Of particular interest here are the two dynamically-inspired claims that many complex behaviours and behaviour developments can be generated without recourse to any internal representations or programs (of both the genetic and neural varieties), and that elements of the agent's surroundings must form a significant part of the causal network that leads to behaviour production.

In what follows I will briefly outline some of the important insights and solutions dynamical theorists have suggested might form the basis of a dynamical framework for understanding cognition and action. I will illustrate these points with examples from research. A more in depth examination of dynamical systems concepts can be found in chapter 5.

Some of the most interesting research conducted under the dynamical systems banner has examined the ways in which complex patterns of behaviour can emerge and evolve from within a body-environment system composed of a collection of coupled subsystems (e.g., link-segment, musculotendon, neural, and circulatory subsystems [Goldfield, 1995, chap.

4]). These systems have occasionally been modelled by dynamical equations that pull out high-level abstract parameters and variables from the behavioural situation. Such research is often couched in terms of the *self-organising* features of systems made up of body, brain, and environmental components. Self-organisation is a characteristic of physical systems that are far from thermodynamic equilibrium (courtesy of the continuous inflow of energy), are dissipative (roughly, creators of disorder), and are complex (composed of a large number of, often heterogeneous, components) (see, e.g., Kauffman, 1993, 1995; Kelso, 1995; Prigogine & Stengers, 1985). Intuitively it may seem likely that such systems would inhabit a massive number of states (have a large number of degrees of freedom) but, under appropriate conditions, they actually spontaneously generate a relatively small number of ordered states. Moreover animals and other living things seem to be fine examples of such systems (see, e.g., Ho, 1993; Kauffman, 1993). A number of researchers are now pursuing the idea that the behavioural patterns exhibited by humans and other animals may be the product of this kind of spontaneous organisation rather than the result of instructional commands sent down by some kind of neurally-realised command and control centre.

These ideas can be illustrated with a widely cited example from the work of Thelen and colleagues (Thelen & Fisher, 1982; Thelen, Fisher, Ridley-Johnson, & Griffin, 1982; see also Thelen & Smith, 1994, chap. 4). They found that the disappearance of infants' reflex stepping pattern around two months of age is due, not to maturational change in some neurally-instantiated central pattern generator, as had been previously hypothesised, but rather to the changes of infants' leg mass due to increases in subcutaneous fat. The stepping reflex disappears simply because the infant does not have the required muscle strength to lift the heavier leg. Thelen and her colleagues showed that this is the case with a couple of clever experiments. In normal non-stepping infants the reflex reappears when the infants are held upright in warm water (thereby reducing the effective mass of the legs) (Thelen et al., 1982), or when they are held upright with their feet touching two moving treadmills (thereby assisting the infant to help stretch their legs backward - the hard part of the task - and letting them complete the stepping pattern with the still controllable forward leg-spring) (Thelen, 1986). Similarly, by increasing the leg mass of stepping infants (by adding small weights) Thelen, et al. (1982) could make the reflex disappear. Thus it appears that the ability to step is an ability that spontaneously self-organises from a system made up of coupled components of leg muscles, leg 'weight' (subcutaneous fat and bone), and environmental support (effective gravity, external leg moving forces). Thelen and Smith (1994, chap. 9) have also used this framework to explore the development of infants' reaching abilities (Thelen et al., 1993) and the development of object permanence skills (measured by the A-not-B problem) (Smith, Thelen, Titzer, & McLin, 1999; Thelen & Smith, 1994, chap. 10), as well as infants' ability to understand the physical laws of objects that Spelke (1990) and Baillargeon (1986, 1987; Baillargeon, Spelke, &

Wasserman, 1985) attribute to innate internal programs (Thelen & Smith, 1994, pp. 223-236).

This kind of framework is not only relevant to the study of motor development in children. One could extend these ideas to explain culturally different walking styles and styles of 'sitting' (amongst other bodily skills). Ingold (1996) reviews literature that points to the different ways in which people of different societies use their bodies to accomplish tasks. Prior to Western influence, Japanese children learnt to 'walk from the knees' rather than 'walk from the hips'. This technique may seem odd to Westerners but, as Kawada (1996 cited in Ingold, 1996) has shown, the resulting lowered centre of gravity is a very efficient way of carrying heavy weights suspended from long poles on uneven ground - a common practice in pre-Western influenced Japanese society. A similar analysis can be had of the Western notion of sitting on chairs versus the much more widespread practice of squatting. These phenomena show that dynamical systems analyses could be fruitfully extended to the more complex issues where particular kinds of artifacts and tools, and the tasks that are accomplished by their use, lead to the self-organisation of energetically efficient bodily postures and skills in much the same way that a cat, for instance, 'adjusts' its gait to deal with uneven terrain or having a leg in a cast. These sorts of ideas have recently been used in the construction of walking robots. Keijzer (1998a) reports related robotics work by Taga and colleagues (Taga, Yamaguchi, & Shimizu, 1991; Taga, 1994) that successfully simulates walking in unpredictable environmental conditions (e.g., changes in profile of the ground) using a self-organising system made up of three coupled subsystems: a rhythm-generating neural network, a musculo-skeletal subsystem, and the local environment. This architecture is radically different from that used in traditional AI and is most easily understood in dynamical systems terms.

These dynamical analyses of self-organising systems have a couple of major advantages over cognitivist-inspired explanations of the production of behaviour. The first advantage is that dynamical systems theory provides a rather nice solution to the *instruction problem* (Keijzer, 1997, 1998a, 1998b). Recall that this is the problem of how it is that a cognitive system can produce essentially the same distal behaviour in many different circumstances via different proximal behaviours. The second advantage of a dynamical approach is that it can quite naturally account for sudden changes in the behaviour of a system without needing to postulate the influence of a new causal variable. As it turns out, both of these aspects of dynamical systems explanations are simple consequences of systems with a particular mathematical or lawful structure. Self-organising systems tend to be both *stable* and able to *inhabit multiple ordered states*. In order to better make sense of these ideas it will be helpful to use some examples from dynamical analyses of simple rhythmic behaviours.

According to Hooker (1997) (after Smith and Thelen, 1994), cognitivist (or, what he calls, *structuralist*) accounts of behaviour fail to explain two important features of animal behaviour: 1) they do not account for the actual fine-grained variability of real-time performances, and 2) they cannot account for the production of essentially similar performances under variable conditions (see also Keijzer, 1997, 1998b). Each time an animal[50] performs a particular *kind* of activity (chewing on some food, walking across an open space, climbing a tree), it accomplishes it using a unique combination of fine-grained movements. As Hooker (1997, p. 104) notes "in each case there is a relevant sense in which it is the same action being performed and yet in each case our precise inner condition is in some degree unique to that occasion." A cognitivist explanation of a behaviour tends to be couched in terms of the broader, behavioural kinds as being 'represented' within the internal plans for action. The question of how such broad plans are tailored to the specific circumstances so that they produce the appropriate finer-grained movements is never properly dealt with. Such issues form the basis of the problematic instruction problem. Dynamical systems explanations, however, can rather easily account for these sorts of issues. The basic idea can be illustrated by Saltzman's (1995) work with phoneme production.

Saltzman and his colleagues (Saltzman, 1995 and references therein) have modelled the human speech-production system using the tools of dynamical systems theory in order to account for a collection of interesting phenomena related to the production of speech sounds. In particular they have produced a dynamical model that can predict the effects of tugging on the jaw as a person attempts to produce a particular syllable (e.g, /bæz/ or /bæb/). What is typically found when this occurs is that other parts of the speech system rapidly compensate (within 20-30 msec) for environmental perturbations (i.e., jaw tugging) so that the same overall outcome (i.e., a particular speech sound) is maintained. The speech system, then, is a resilient, roughly homeostatic, self-organising system made up of a collection of equally contributing parts that continuously modulate each other's activities. Such a system can be dynamically modelled using a collection of rather abstract variables and parameters that are grounded in the collective activity of the speech system anatomy. This is in contrast to the cognitivist vision of a neurally-instantiated planner and controller

---

[50] Hendriks-Jansen (1996, pp. 141-149) notes that this is also the case for Brooksian mobots. In his discussion of Mataric's (1991) navigating robot he points out that, although the robot can be understood to be 'wall-following' at a broad grain of analysis (what Keijzer [1997, 1998b] calls its *distal behaviour*), no two wall-following events are ever exactly the same at a finer-grain of analysis (what Keijzer [1997, 1998b] calls its *proximal behaviour*) because the behaviour produced is an interactively emergent product of robot and environment. Brooksian mobots do not 'behave' by attempting to execute an internal 'movement template' (i.e., set of prescribed movements). The patterns of activity produced by these robots cannot be summarised as if they were variations on a particular sequence of spatial coordinate transformations.

that 'tells' each peripheral body part what to do – what Saltzman refers to as the strawman hypothesis that "a central executive or homunculus ... produces a given movement pattern with reference to an internal kinematic template of the form, tracing out the form provided by the template, and using the articulators as a physiological and biomechanical pantograph to produce a larger version of the pattern in the external world." (p. 151)[51]. In such a case we would predict that the perturbing of one's subsystem's activity (i.e., the jaw) would disrupt the phoneme production by altering the overall shape of the speech system. (Saltzman, 1995, p. 158).

Thus the dynamical model not only provides accurate predictions of such behaviour in a way the representational-computational model usually does not, but it also can account for the common intuition that "there is a subtle underlying invariance of control despite an obvious surface variability in performance." (Saltzman, 1995, p. 152). In sum, a dynamical description of a self-organising system nicely accounts for the ways in which the system compensates for environmental peculiarities in the production of a particular kind of behaviour. But such systems are not doomed to a kind of 'inflexible' homeostasis where every environmental change is compensated for by the system's self-organising dynamics. Obviously no sensible animal tries to do the same thing in the face of every kind of environmental perturbation. The very same sort of system that exhibits stability over a range of environmental perturbations can, seemingly suddenly, jump to a new ordered state when a certain 'intensity level' of environmental 'influence' is reached.

This is important for dynamical explanations of behaviour and cognition because it suggests a mechanism whereby a single dynamical system can exhibit a reasonably stable set of behavioural responses within a certain range of environmental conditions as well as an ability to jump into a new adaptive behavioural state once a certain threshold is reached. For example, such a dynamical system may be at work in an animal that can continuously maintain a safe distance from a potentially dangerous rival that is moving within a range of relatively slow speeds, but that turns tail and runs when the rival's speed reaches a certain critical point.

Such phase transitions have been investigated and subsequently modelled by Kelso and his colleagues' study of rhythmic behaviours such as walking, swinging arms and (in one widely cited study) finger waggling (see Kelso, 1995 and references therein). Kelso and his colleagues found that people can waggle their two index fingers side to side either in-phase

---

[51] Although this view is indeed a strawman in its baldest form, the general idea is alive and well within modern cognitive science. Pinker (1997, pp. 11-12), for instance, compares a human arm to an architect's lamp and marvels at the complex computations it must perform to get the trigonometry right and to compensate for the effects of the arm's momentum as it moves about.

(moving toward each other and then away from each other) or anti-phase (moving like windshield wipers on a car) but not (easily) at any other phase when the frequency of waggling was relatively low. These two states can be maintained over a range of frequencies but, at a certain higher frequency the system becomes unstable and suddenly kicks into a single state in-phase regime[52]. It is likely that such a change occurs because, at certain higher frequencies, there exists only one state that is energetically efficient[53]. The dynamics of this phenomenon have been successfully modelled by Haken, Kelso, and Bunz (1985 cited in Kelso, 1995) with their HKB model. This equation includes the control parameter of 'waggling frequency' and the collective variable *relative phase*. With it the change from one attractor dynamics to another can be accurately predicted and, indeed, explained (see chapter 5 for the discussion of the vexed question of whether dynamical analyses actually *explain* behaviour). This model successfully explains many kinds of interlimb coordination and, with a few modifications may provide the basis for a *general theory of meter* which may be applicable to other fields such as simple speech tasks (Chemero, 1999, pp. 22-23).

These kinds of phenomena are nicely illustrated by dynamical analyses of the gait changes in quadrupeds and, more recently, in hexapods such as the humble cockroach (see, e.g., Beer, 1995a, 1995b; Stewart, 1998, chap. 9; Turvey, 1990). The research in this area of interlimb coordination shows that changes in gait (from trotting to cantering to galloping and so on) occur as the rate of movement (the control parameter) increases. The savings in energy conferred by such changes are obviously beneficial (adaptive) for fleeing or chasing animals (i.e., they can go further and faster by doing this). These sorts of changes have been modelled using relatively uncomplicated dynamical analyses as well as using dynamical interpretations of pattern-generating neural networks. Such modelling has shown that a single neural pattern generator with a nonlinear dynamics can account for the kinds of gait pattern that emerge at different movement rates. The need to embrace the more cognitivist ploy of multiple central pattern generators or a single pattern generator with multiple 'gait programs' is refuted by such research (see Thelen & Smith, 1994).

---

[52] In the language of the dynamical systems equations used by Kelso and his colleagues, the in-phase condition reflects an attractor at relative phase 0 and the anti-phase condition an attractor at relative phase .5. There is a change in the phase portrait (a phase transition) at the critical frequency point which results in the disappearance of the relative phase .5 attractor.

[53] Thelen (1994, p. 342) notes that attractor states may also result from "special learning". By this she presumably means movement patterns that people and animals 'fall into' as the result of intensive training. For instance, Stewart (1998, p. 178) notes that most horses do not naturally canter (attractors at relative phase 0, .8, .8, and .5 for the front left, front right, back left, and back right legs) but must be trained to do so.

In sum, dynamical systems researchers have provided arguments and evidence to suggest that behaviour can be generated in distinctly uncognitivist ways. First, stable, ordered patterns can be produced by far from equilibrium systems made up of coupled components. In such systems potentially massive problems of controlling the possible states that an agent may enter are reduced by the mutually constraining effects of the coupled system components.

Second, such systems require little in the way of a 'command and control' centre for computing how an animal or robot should move its body in order to accomplish a particular kind of behaviour. The neural component of such a behaviour system may consist of no more than a rhythm generator or oscillator which supplies vital but insufficient structure (or 'information') for producing the appropriate behaviour. Other subsystems, including the local environment, are equally important 'informational' partners. Such observations can be used as ammunition against cognitivist arguments that complex behaviour necessarily requires the manipulation of complex internal representations of the environment that are had via the transduction of energy into neurally-coded semantic information.

Finally, self-organising systems are often resilient in the face of environmental perturbations. These systems can often produce 'task appropriate' behaviour in a variety of conditions via the rapid compensation by other subsystems. In addition changes in environmental conditions may 'push' the system into a new stable behavioural pattern more suited to dealing with the current environmental climate.

Cognition in the Wild: Distributed Cognition and Situated Action Research

The past decade and a half has witnessed a move amongst some researchers to study human cognition in the full cultural and natural complexity of people's everyday worlds – a move which Hutchins (1995a) has aptly referred to as the study of 'cognition in the wild' and Geertz (1983) has called 'outdoor psychology'. Typically the psychologists, sociologists, and anthropologists engaged in these studies do not consider themselves to be doing applied research but rather ecologically valid pure research into the fundamentals of cognitive activity. They have studied cognitive activity in situations such as photocopier use (Suchman, 1987), nuclear powerplant operation (Vicente & Burns, 1996), maritime navigation and piloting (Hutchins, 1995a), aircraft piloting (Hutchins, 1995b), formal schooling programmes (Greeno & the Middle School Mathematics Through Applications Project Group, 1998; Perkins, 1993), work in a milk-processing plant (Scribner, 1986), the organisation of children's mealtimes (Valsiner, 1997), and everyday uses of mathematics in shopping, budgeting, and dieting (Lave, 1988).

I have grouped together distributed cognition and situated action research because the two frameworks share many similar ideas about cognitive activity. It should be noted, however, that situated action research has a different flavour to it from distributed cognition research and that this flavour derives from its different intellectual background. Distributed cognition research owes an allegiance to some basic sort of information-processing account. Situated action research, by contrast, derives from sociological and anthropological movements such as ethnomethodology (Suchman, 1987, 1993) and practice theory (Lave, 1988), as well as ecological psychology (Costall & Leudar, 1996; Vicente & Burns, 1996). It is probably fair to say that most distributed cognition and situated action researchers are aware of, and sympathetic to, the ideas in the other related fields and differ primarily in their emphases and reasons for engaging in research[54]. Of course, even within each programme one finds different ways of doing things and interpreting phenomena. My emphasis here will primarily be on the similarities. Where appropriate I will signal which tradition provides a more in depth analysis of a particular aspect of the shared problem domain. The sketch I provide here draws heavily on the work of the cognitive ethnographer Edwin Hutchins (1995a, 1995b) but it is also informed by the theorising of Norman (1993a, 1993b), Pea (1993), Perkins (1993), Suchman (1987, 1993) and others.

The Distribution of Resources for Cognitive Activity

The fundamental assumption underlying distributed cognition research is that most everyday cognitive activity cannot be understood as purely in-the-head cogitation, because it takes place within a wider system that includes resources from the local physical and social environments – environments, moreover, that are usually themselves the ever-changing product of cultural evolution. Indeed the environments in which even the most mundane social activities occur are in an important sense 'designed' to make certain cognitive activities possible. Thus distributed cognition researchers often endorse a softening of the boundary between brain and world in their analysis of cognitive activity. The furniture of our surroundings actually forms part of a distributed cognitive system that Hutchins (1995a) calls a *functional system* rather than merely providing inputs and a place to 'output' behaviours. The structured surround is viewed as providing important cognitive resources – vehicles for thought, stores of information, control structures, as well as goal-setting and decision-making systems. Pea (1993) claims that "the distributed-intelligence framework sees a much more substantial haze around the boundary of the person and shines the light of attention on the more invisible intelligence in the artifactual, physical,

---

[54] See the special issue of *Cognitive Science* on Situated Action (Norman, 1993a) for a fairly thorough examination of the varied ideas held by proponents of the different approaches.

symbolic, and social surrounds, as brought into relief in the configurations of distributed intelligence by which activity is achieved." (pp. 53-54).

Thus we find, for instance, Perkins (1993) advancing what he calls, the *equivalent access hypothesis.* Roughly speaking, this is the claim that the locus of knowledge 'storage' and knowledge construction in a 'knowledge-processing-system' is not always in the agent's head, and that the functionally important aspect of such a system is *the degree and efficiency of access to knowledge* vital for the performance of a particular cognitive activity. Perkins, thus discusses distributed cognition within an *access framework.* Like Hutchins he treats external mediating structures as if they were 'parts' of 'the mind' and uses a modified information-processing model to describe their properties and processes[55] (see also Clark & Chalmers, 1995). However, Hutchins (1995a) has moved well beyond a simple application of information-processing concepts to broader functional systems and has introduced a variety of new terms and concepts for making sense of the unique phenomena associated with cognition within distributed cognitive systems. I discuss many of these ideas in the following sections.

Functional Systems and Mediating Structures

Extended cognitive systems are understood as *functional* systems because the focus is on the functions of parts of the cognitive system rather than whether those parts are located inside or outside the skull. A functional system is a dynamic entity made up of continually changing interacting *mediating structures* (a.k.a. *media* or *structures*). Mediating structures are simply tools or artifacts or human cognitive abilities that can both *hold* information in a particular format and *transform* that information in a useful manner. One of Hutchins' (1995a) favourite examples of a mediating structure is the nautical slide rule. Others include calculators, sketch pads, maps, physical models, filing systems and so on. In functional systems information is propagated across the various mediating structures until it reaches a point where it can be used to produce some kind of action. By *propagation* Hutchins (1995a) means that the transformed information from one mediating structure is fed into a subsequent structure where it can be further manipulated. These ideas are nicely illustrated in Hutchins' study of the dynamics of a US Navy navigation team.

Hutchins (1995a) focuses his attention on the work done by a navigation team on the bridge of the military vessel (an amphibious helicopter transport) that he calls the *USS*

---

[55] In Hutchins' case the use of information-processing terminology is explicitly metaphorical. At times Hutchins suggests that using the information-processing framework to understand 'high-level' socio-technical systems seems "a much more solidly grounded application of the computational metaphor to a cognitive system than the application of this metaphor to the workings of an individual mind." (1995a, p.185, see also chap. 9)

Palau (a pseudonym). His particular concern is with the way in which the ten-member team[56], known as *Sea and Anchor Detail*, takes bearings in order to keep a constant record of the ship's location as it traverses shipping lanes near harbour. The position fixing activity performed by Sea and Anchor Detail is known as a *fix cycle*. It consists of the following operations and interactions (see Hutchins, 1995a, chap. 3, 4, and 8 for more detail).

First, a seaman (the *bearing taker* or *pelorus operator*) stationed on either the port or starboard bridge wing points his *alidade* (a kind of combination telescope and compass readout) at a designated landmark and takes a reading from the gyrocompass scale. This turns a visual perspective into a string of three digits that represent the actual bearing of the landmark with respect to true north (the "true bearing"). The pelorus operator communicates this number to the *bearing recorder* over a phone circuit. The recorder notes the figure in the bearing log in standard digital format. He then communicates this information to the *bearing plotter,* who stands near the recorder in the pilothouse. The plotter uses the spoken information to set his hoey (a protractor with a long 'arm' attached). The spoken bearing digits are used to set the hoey's arm at a particular angle. The hoey is then brought into coordination with the chart of the local waters. The angle of the hoey's arm is used to mark the chart with the ship's *line of position* with respect to the landmark originally spotted by the pelorus operator. A line of position (LOP) is a line that passes through both the ship and the landmark. The LOP does not fix the exact position of the ship on the chart. Rather the vessel lies somewhere upon that line. In order to fix the ship's position another two LOPs must be plotted using two more landmarks. The area (a triangle) about which the three lines intersect represents the ship's approximate position.

In the navigation team's attempts to plot fixes of the ship's position, information necessary for the plot-fixing task enters the system with the pelorus operators finding and 'shooting' a designated landmark with their alidades. Information is then grasped from this situation and embedded in various artifacts in different representational formats. Each format is tailored to different subtasks. Digital format can serve to make arithmetic operations possible or to enable 'portability' and transmission over the limited bandwidth of the phone circuit (i.e., its easier to say a number than to try and describe the relationships in the visual scene). Analog format, such as the angle of the hoey's arm and the layout of the nautical chart, makes it possible to carry out geometrical computations such as plotting the ship's position on the chart.

---

[56] See Hutchins (1995a, pp. 178-185) for a breakdown of the different crew in this team. Here I only focus upon the two pelorus operators, the recorder, and the plotter. Other crew include helmsmen and the fathometer operator.

When this cycle is performed time after time the position of the vessel can be seen as the point where multiple lines of position intersect. Hutchins' analysis of various fix cycle events on the Palau is fascinating because it reveals a number of possibly counter-intuitive issues associated with real-world cognitive activity. Firstly, he shows that only by examining an entire functional system (in this case the navigation team, their tools, practices, and accompanying social positions) can an accurate picture be built up of what is going on and why. Elsewhere Hutchins argues forcefully that "the outcomes of interest are not determined entirely by the information processing properties of individuals. Nor can they be inferred from the properties of the individual agents, alone, no matter how detailed the knowledge of the properties of those individuals may be." (p. 265).

So in Sea and Anchor detail the task of plotting the ship's position is achieved by a functional system that includes at least four crew members, alidades, a gyro-compass, phone circuit, bearing log, hoey, chart, at times a calculator, as well as the social rules and routines and normative computational procedures of naval practices. The behaviour of interest, the position-fixing, is not controlled or fully comprehended by any individual (and thus is not explainable in terms of any crew member's cognition). Indeed, many of the computations that occur within the task do not occur in any one's 'head' but are simply made by manipulating a tool or artifact and passing the results of that manipulation onto the next tool in line.

In order to discuss and theorise about the distribution of cognitive process implied by cognitive activities such as the routines of the Sea and Anchor Detail, Hutchins has coined a number of useful terms for understanding the roles of artifacts and other mediating structures in distributed cognitive activity. Mediating structures can be understood to contain content-level knowledge, contain complex tacit knowledge, distribute workload over time, constrain the organisation of activity, and transform the kinds of activities that people perform.

Mediating Structures Contain Content-Level Knowledge

Mediating structures such as artifacts can contain what Perkins (1993) calls *content-level knowledge* in a format specially tuned to its place within the broader *cognitive ecology* of the functional system. Content-level knowledge is simply that basic sort of information used in a task that can often be 'kept in your head' such as a compass bearing or phone number. Tables, databases, books and so on are typical repositories of content-level knowledge. The notion of *cognitive ecology* refers to ways in which the mediating structures (including artifacts, techniques etc.) fit and relate to each other. Particular artifacts or techniques are usually designed so that their products (i.e., output) mesh in some useful way with other mediating structures within the system.

Mediating Structures Contain Complex Tacit Knowledge

Artifacts (and other media) also 'store knowledge' in a deeper sense that includes knowledge of what to do and how to do it, and when it should be done. Hutchins (1995a) refers to this knowledge as *artifact memory*. Artifacts typically 'contain' a kind of *tacit knowledge about procedures or things* that would be difficult or impossible to use if one was relying entirely upon one's native cognitive abilities. Tacit knowledge is knowledge that *cannot* be simply read off an artifact, such as a bearing on a compass. Rather it is 'knowledge' in the sense of structural relations and regularities of the artifact that constrain the kinds of calculations or cognitive products that can be generated using the artifact. Hutchins (1995a) argues that these action constraints typically "embody kinds of knowledge that would be exceedingly difficult to represent mentally." (p. 96). He illustrates these ideas by examining the astrolabe, an ancient navigational device for predicting the movement of the sun and the stars at various latitudes. The astrolabe can control spatial relationships of the heavens and is precise and durable in a way that one's own 'mental models' could never be. The astrolabe's tacit knowledge exists in its ability to maintain and transform the spatial relationships between many heavenly bodies. Such knowledge simply be 'read off' and used subsequently in unaided mental modelling.

Mediating Structures Distribute Workload Over Time

Importantly, artifacts can make activities possible and/or alter the time they take or the way they are done by *distributing the workload involved in cognitive activity over time* – a phenomenon that Hutchins refers to as *precomputation*. Precomputation is simply the preparation of the working environment in advance so as to support cognitive activity when a particular task arises. Hutchins (1995a, pp. 167-168) notes that we should understand precomputation as occurring on various time-scales from the "setting of the hoey, which was done a few seconds ago" to the "changes to the chart that were plotted a few days ago" to the "nature of the plotting tools, which were designed a few decades ago" to the "mathematics of the projection chart, which was worked out a few centuries ago" (pp. 167-168). Each of these past cognitive achievements plays an important role in the production of the present activity. A full and proper understanding of a cognitive activity requires us to acknowledge the ways in which a person, and others inhabiting the person's environment, have modified tools, artifacts, and layouts of the environment in order to change (improve, speed up, make possible) the ways we perform those cognitive activities. Precomputation is made possible by 'freezing' task-specific invariants (things about the task that we are doing that do not change over a significant period of time) into our tools and layouts (Hutchins, 1995a, pp. 165-167). Information that does not change over a period when a task may be carried out can be 'frozen into' an artifact or a layout so that a user does not need to needlessly re-calculate that required information every time the task is performed. A simple example may include stacking nautical charts in the order in which

they will need to be used. This cuts down on the time needed to search through chart piles. Indeed the ability to 'search' is frozen into the environmental layout so that the simple situated skill of 'moving the top chart' will instantly 'locate' the next required chart. A more complex example can be seen in the use of the parallel motion protractor (PMP). When used in the navigation the PMP can automatically correct bearings for magnetic variation. Without this ability one would have to plot a line of position and then modify it with information about the difference between true bearings and magnetic bearings. Thus, by freezing this invariant (which is invariant over tens of nautical miles and tens of years) into the structure of the PMP, navigators relieve themselves of the task of doing an extra 'reflective' cognitive calculation.

Mediating Structures Constrain the Organisation of Activity

Artifacts also serve as *constraints on the organisation of action* by having particular ways of doing things embedded into their structure. Artifact-filled functional systems can increase the efficiency and decrease the error-rate of cognitive activities by making it impossible to perform incorrect calculations. This is achieved by building error-proof constraints into the artifacts or layouts themselves. The nautical slide rule (and its relative the nomogram) prevent operations that violate the rules (syntax) of nautical calculations regarding distance, elapsed time, and rate of speed. They "obviate or lock out such relations among terms. The relations $D=RT$, $R=DT$, and $T=D/R$ are built into the structure of the nomogram and slide rule. The task performer has no need to know anything about these relations, either implicitly or explicitly." (Hutchins, 1995a, p. 150). Thus, artifacts can guide the user to correct conclusions in various ways by, in a sense, setting goals and making decisions. By constraining the kinds of possible outcome, the slide rule is 'deciding' what kind of operation is permissible. By contrast an individual performing speed calculations may arrive at an incorrect result by, for example, deploying the wrong formulae (e.g., $D=T/R$).

Mediating Structures Transform the Cognitive Activities Required of the Agent

The use of an artifact therefore *transforms the cognitive activities* that the navigator, in this case, needs to carry out to accomplish the task at hand (Hutchins, 1995a, pp. 153-155, p. 170). Central to the notion of precomputation is the fact that precomputing usually changes the component activities required to accomplish a particular task. The building of a calculator changes addition from a pencil and paper or in-the-head procedure for implementing simple algorithms to a button pushing, screen-reading activity. The PMP changes the calculation of magnetic variation from a computation to a simple scale-alignment and scale-reading operation.

Both the distributed cognition and situated action frameworks take the above focus reliance on the environmental backdrop for activity as signalling a new view of what it is that the

individual does and contributes to cognitive activity. Cognitivism and associated philosophies in the social sciences typically take the individual to be a relatively self-contained planner of activity and controller of the world. By contrast both distributed cognition research and situated action research paint a picture of the agent as mostly in the job of coordinating external cognitive resources and using them to promote the use of more embodied, perception and action skills. For instance, instead of engaging in *effortful mental arithmetic* (internal cogitation) to evenly divide a pile of objects into several smaller piles, a person might just sort objects into piles that *look* to be about the same size (effortless, embodied, perceptual skill). Within the navigation context this idea is nicely illustrated by the way in which a nautical slide rule can be used to transform the effortful activity of doing mental arithmetic (multiplication, division, calculating distance covered from current speed and time, etc.) by simply aligning the scales on the slide rule using simpler perception and manipulation skills. The 'difficult' cognitive activities of reasoning and problem solving are often accomplished, not so much by individual, internal cogitation, as by rearranging tools and other environmental entities in order to present us with simpler and more 'intuitive' ways of getting information from the world. We use new skills to complete tasks; often swift, simple, 'situated' or 'experiential' skills such as object manipulation and pattern recognition, to perform what would otherwise be time-consuming, effortful (or even impossible) 'reflective' skills.

> These tools permit us to transform difficult tasks into ones that can be done by pattern-matching, by the manipulation of simple physical systems, or by mental simulations of simple physical systems. These tools are useful precisely because the cognitive processes required to manipulate them are not the computational processes accomplished by their manipulation. The computational constraints of the problem have been built into the physical structure of the tools (Hutchins, 1995a, p. 171).

Norman (1993b, pp. 26-27) notes that these situated and experiential procedures, that are deployed when cognizing with the aid of artifacts, are often less effortful and prone to disruption than those that require reflection and concentration. Precomputation can save time and effort and avoid error by transforming the type of problems people need to tackle.

Thus the basic idea that underlies much of the theorising in distributed cognition and especially situated action research is that people are 'designed' to make use of mediating structures. Human cognition is intrinsically *cultural* in the sense that people make and use environments that promote and support particular kinds of activity (Hutchins, 1995a). Moreover this process is an importantly cultural and historical feature of human evolution[57]. People have systematically modified their environments in order to "... change

---

[57] It is certainly within the realms of possibility to argue that human anatomy and physiology have co-evolved with certain cultural practices. Deacon (1997), for instance, argues that language has co-evolved with human culture, over millions of years. And, if we take notice of the claims of evolutionary

the nature of certain computational problems so as to make them more tractable to perceptual, pattern-completing brains ..." (Clark, 1997, p. 77) for thousands if not millions of years. As we have seen, Hutchins (1995a) makes this clear when he argues that we must understand much modern cognitive activity in the context of, often hundreds of years, of precomputing by our ancestors. What we do now relies on what has been done before both by ourselves and especially by others. Human cognitive activity is deeply and intrinsically embedded in the ways in which people have systematically altered their surroundings.

The vision of the agent that emerges from this analysis is one where our primary engagement with our surroundings is at the level of pattern recognition and skilful sensorimotor activity. Distributed cognition and situated action analyses seem to strip the individual of the traditional cognitivist powers of reflective cogitative, creativity, and effortful planning and replace them with a manipulator of 'mediating structures' and a recogniser of patterns. The typical agent becomes more of a 'situated seer and doer' (a specialist in experiential cognition [Norman, 1993b]) deploying embodied skills rather than behaving like a Rodin-like thinker.

Summary

It may seem that distributed cognition and situated action research focuses on an entirely different level of analysis compared to that of basic cognitive psychology. The focus on external resources in the production of cognitive activity may appear to be simply an attempt to apply basic cognitivist analyses to real-world situations. Indeed much distributed cognition and situated action research has started with this possibility in mind. What distributed cognition and situated action researchers have found, however, is that the traditional cognitive psychology approach of thinking that most cognitive activity is an internal affair that involves some kind of manipulation of inner models in order to produce plans for action conflicts with the stubbornly 'situated' approach people seem to take to many cognitive tasks; "[i]t is notable how vigorously we human beings, given half a chance, function as agents recruiting into the cognitive enterprise not only other people but the insentient physical things around us, arranging them and refashioning them so that they become 'partners in cognition'." (Perkins, 1993, p. 107).

---

psychologists like Barkow, Cosmides, and Tooby (1992) who argue that modern human biology and psychology is adapted to our hunting and foraging niche in a Pleistocene environment, it makes sense to suggest that we are likely to have, for instance, neural traits that support living and reproducing in Pleistocene *societies*, which involved the use of tools, artifacts, and various social practices. Cosmides (1989), for instance, has argued that we possess a cheater detection 'module' to cope with such a social environment. It is only a short step from such claims to the idea we may well have co-evolved situated cognitive mechanisms and the ability to produce mediating structures.

Everyday human cognitive activity seems to be defined by the fact that people lean heavily on modifications they, or others, have made to the environment and its furniture so that they can deploy fast and often efficient experiential cognitive abilities such as pattern recognition and sensorimotor skills. Theorists like Suchman (1987, 1993) and Costall and Leudar (1996) argue that modelling and planning, even in its so-called subpersonal and unconscious guise, is not as ubiquitous a feature of thoughtful human activity as traditional cognitive science has claimed. When they *do* occur such processes should be understood as particular kinds of personal-level *activity* that are products of social/cultural learning, rather than being steps in every internal computational implementation of a cognitive activity (see also Clancey, 1993, 1997). Distributed cognition and situated action researchers argue that traditional cognitive science has misrepresented everyday human cognition in an important way by making it seem a good deal more cogitative and internally focused than it really is. By contrast distributed cognition researchers such as Hutchins (1995a)

> believe that the real power of human cognition lies in our ability to flexibly construct functional systems that accomplish our goals by bringing bits of structure into coordination. That culturally constituted settings are rich in precisely the kinds of artifactual and social interactional resources that can be appropriated by such functional systems is a central truth about human cognition. The processes that create these settings are as much a part of human cognition as the processes that exploit them, and a proper understanding of human cognition must acknowledge the continual dynamic interconnectivity of functional elements inside with functional elements outside the boundary of the skin (p. 316).

Critics such as Clark (1996b) argue that we must be careful how we use these distributed cognition and situated action arguments for it is surely the case that people can, and often do, 'do things in their heads' without relying upon external cognitive resources. This, Clark argues, leaves room for a modified computational view of mind, one he believes will be based upon something like the architectures of artificial neural networks (see Clark, 1993, 1997). Clark is, without doubt, correct in making this observation and I suspect that most distributed cognition and situated action theorists would agree with the spirit of his criticism. The question is, however, just what kind of 'internal thinking' will fit the bill and how much of it comes to us naturally and how much is a product of 'internalising' skills, such as mental arithmetic, that we originally learnt in an extended cognitive system consisting of such things as teachers, pencils, paper, and number blocks. The notion that private, in-the-head thinking may arise from the appropriation of public, social skills is central to the sociohistorical approach pioneered by Vygotsky. It is to this perspective that we turn next.

The Sociohistorical Approach

The sociohistorical (a.k.a. sociocultural, cultural-historical, Vygotskian) approach to the study of development derives from the work of Vygotsky (1930-1935/1978, 1934/1986)

and his colleagues and students including Luria and Leontiev[58]. Although the sociohistorical approach is often only mentioned in passing, if at all, in most modern discussion of ESD perspectives on cognition, I believe that it contains a number of important insights for interactionist theorists.

Vygotsky's work has been developed in a number of directions by psychologists working in several countries. Vygotsky's sociohistorical ideas are important historical precursors to many of the ideas within distributed cognition research and, to a lesser extent, situated action research. Vygotsky, like distributed cognition and situated action researchers, emphasised the role of resources outside of the individual's 'natural mind' in the production of complex human cognitive activity. In particular, Vygotsky and his colleagues focused upon the ways in which humans make use of both *technical tools* (e.g., spades, cars, spectacles) and *psychological tools* (e.g., words, routines, books, the assistance of an expert) to *mediate* their interactions with their surroundings.

> Modern man does not have to adapt to the external environment in the way that an animal or primitive man does. Modern man has conquered nature and what primitive man did with his legs or hands, his eyes or ears, the modern man does with his tools. Cultural man does not have to strain his vision to see a distant object – he can do it with the help of eyeglasses, binoculars, or a telescope; he does not have to lend an attentive ear to a distant source, run for his life to bring news, – he performs all these functions with the help of those tools and means of communication and transportation that fulfil his will. All the artificial tools, the entire cultural environment, serve to 'expand our senses' (Viner, 1909). Modern cultural man can allow himself the luxury of having the worst natural abilities, which he amplifies with artificial devices thus coping with the external world better than the primitive man who used his natural abilities directly. The latter broke a tree by beating it on a stone, modern man takes an ax or a frame-saw and does this work quicker, better, and with less energy wasted. (Vygotsky & Luria, 1993, pp. 169-170)

Aside from its old-fashioned, progressivist ideas the framework set out here is deeply reminiscent of those described by Hutchins (1995a) and Perkins (1993) in the previous section. The framework sketched by Vygotsky and his contemporaries has been fleshed out in different ways by a variety of modern theorists and researchers (e.g., Berk, 1994; Bruner, Greenfield, & Olver, 1966; Scribner & Cole, 1981; Rogoff, 1990; Nelson, 1996; Valsiner, 1997; Wertsch, 1985, 1991).

However, this emphasis on the distribution of cognitive resources formed only one thread in Vygotsky's work. Vygotsky's other focus involved the hypothesis that psychological tools could gradually be taken on board by the individual and used to restructure their basic mental abilities. Indeed, Vygotsky saw this sort of cultural process as central to the development of modern humans. Vygotskian socio-historical views oppose or complement (depending on one's theoretical predilection) the individualistic and self-organising ideas

---

[58] See Kozulin (1986) for a more complete list of Vygotsky's collaborators and their contributions.

of Piagetian approaches to development. Instead of following Piaget's (1970) vision of the child as moving from an egocentric being to a socialised one, sociohistoricists argue that children first learn how to live and act in the social world and then progressively *internalise* (or *appropriate*) these intrinsically social resources for their own individual uses. The individual skills of reflective thought, voluntary memory, and so on, are the products of social activity rather than necessary components for making that activity possible. Early childhood consists of a gradual taking over of the behaviour-regulating activities and processes that parents and caretakers use to direct and focus the child's behaviour. This basic idea is given in Vygotsky's oft quoted claim that:

> Each function in the child's cultural development appears twice: first, on the social level, and later, on the individual level; first, between people (interpsychological), and then inside the child (intrapsychological). (Vygotsky, 1930-1935/1978, p. 57)[59]

A well rehearsed example of this principle comes from Vygotsky's analysis of the emergence of pointing in infancy from a grasping action toward an out of reach object (see Kozulin, 1986, p. xxvii; Lock, 1980). Initially this grasping action (that Vygotsky referred to as the action-in-itself) is nothing more to the child than a personal attempt to reach something. However, an adult or more socially experienced partner may perceive the action as a pointing gesture, or, in any event, see it as an attempt by the child to get an object. This person naturally helps the child to get the object (or if the object is 'off limits', move it from their reach). Either way the object enters into a triadic relationship with the child and the adult. The network as a whole 'interprets' the action as a pointing gesture. Vygotsky called this state of affairs the 'action-for-others'. The child's experience with such situations leads to a gradual bringing under control and tuning of the grasping movements by the infant as they become a socially communicative act of pointing complete with a dawning awareness of the meaning (social implications) of the action. Deliberate pointing becomes, what Vygotsky called, an 'action-for-oneself'.

This movement from public to private cognitive ability is particularly important in the development of both thought and language. Thinking and speech are seen as independent cognitive abilities that come together, interpenetrate each other, and then develop reciprocally over the lifespan. Particular emphasis is given to the role of public symbolic ability (speech, the use of mnemonics, and memory aids, etc.) in the restructuring of basic cognitive abilities (the natural or elementary mental functions such as elementary perception, memory, attention, and will [Kozulin, 1986, p. xxv]) into "higher mental functions" such as voluntary memory, reflective verbal thought and the like. The basic idea is that the speech of others is used to influence the child ("take this", "don't touch that!")

---

[59] Kozulin (1986, p. xxvi) credits Pierre Janet with being the first to propose this kind of idea.

and, as the child comes to learn how to use language to communicate and 'regulate' the behaviour of others, they find that they can redirect that speech toward themselves to remind themselves of things and to focus their attention (Berk, 1994). Thus, such abilities are evidenced by an increased capacity for the self-regulation of behaviour (the ability to work and think independently of the external constraints of social norms and physical prompts, restraints, and environments). The following passage from Vygotsky (1930-1935/1978) shows how he applies these ideas to the development of voluntary memory:

> A comparative investigation of human memory reveals that, even at the earliest stages of social development, there are two, principally different, types of memory. One, dominating in the behaviour of nonliterate people, is characterised by the non-mediated impression of materials, by the retention of actual experiences as the basis of mnemonic (memory) traces. We call this natural memory, and it is clearly illustrated in E. R. Jaensch's studies of eidetic imagery. This kind of memory is very close to perception, because it arises out of the direct influence of external stimuli upon human beings. From the point of view of structure, the entire process is characterised by a quality of immediacy.
>
> Natural memory is not the only kind of memory, however, even in the case of non-literate men and women. On the contrary, other types of memory belonging to a completely different developmental line coexist with natural memory. The use of notched sticks and knots, the beginnings of writing and simple memory aids all demonstrate that even at early stages of historical development humans went beyond the limits of the psychological functions given to them by nature and proceeded to a new culturally elaborated organization of their behaviour. (Vygotsky, 1930-1935/1978, p. 38-39)[60]

In the terminology used here, sociohistorical approaches view our distinctly human time-space distancing powers as arising from the immersion of the relatively 'naked brains' of infants in complex social and cultural environments.

A classic example of this principle is supposedly evident in an experiment conducted by Alexei Leontiev in 1932 (Kozulin, 1986, p. xxviii). In this study Leontiev supplied a set of coloured cards to people from three different age groups (preschoolers, adolescents, and adults) to use as external support in a colour naming game. In this exercise the participants were required to answer the experimenter's questions without using certain forbidden colour names. The cards could therefore be used as reminders (psychological tools) of the forbidden colours and consulted as each question was asked. Leontiev found that success increased with age. More significantly, however, he found that the preschoolers and the adults did not use the cards as external memory and attention scaffolds. The adolescents, on the other hand, did. Leontiev, and Vygotsky after him, argued that the preschoolers had

---

[60] This distinction is reminiscent of Tulving's (1972) distinction between semantic and episodic memory. Sherry and Schacter (1987) provide a non-social, evolutionary hypothesis for the existence of these two kinds of memory. Nelson (1993, 1996) combines evolutionary and social ideas by arguing that semantic (generic memory) and episodic memory exist in some other higher animals and thus are hard-wired into

not yet learnt how to mediate their natural cognitive abilities using symbolic structures whereas adolescents and adults had. Adolescents by using the actual cards and adults by using a similar 'internal' strategy. They thus concluded that this evidence supports the contention that people gradually develop an ability to mediate their natural (elementary) abilities using socially-supplied symbolic structures (whether externally or internally) to produce higher psychological abilities. With development, skills for using external symbolic structures (psychological tools) to mediate activity are *internalised* as, for example, inner speech and mental imagery.

Leontiev's demonstrations may not convince the modern psychologist who may argue for different interpretations of the data and perhaps question the accuracy of the data in the first place[61]. However a good deal of recent research points to similar conclusions albeit in slightly modified form (see, e.g., the discussion of private speech research in the next section).

Although these laboratory experiments were suggestive, Vygotsky's primary interest was in the ways different cultural practices, such as literacy, impacted upon and reorganised people's cognitive abilities. Vygotsky arrived at similar conclusions to those made from the laboratory studies from his expedition with Luria into the rapidly industrialising parts of Soviet Central Asia in the 1930s (Luria, 1976). They found some surprising differences in the cognitive skills of illiterate peasants and their recently formally educated kin. For instance, they found that those who had not been exposed to Soviet education had extreme difficulty (almost an unwillingness) in solving abstract reasoning problems that conflicted with concrete, sensible everyday situations – the sorts of reasoning skill that Piaget claimed nearly everyone develops when they move to the formal-operational stage)[62]. Vygotsky

---

humans. Autobiographical memory (a key part of episodic memory), on the other hand, arises from social influences on episodic memory.

[61] Where, for instance, are the checks to see whether the preschoolers understood the task in the same way as the adolescents and the adults? It also seems questionable that adolescents would be unable to 'imagine' using the coloured cards in the same way that the adults were supposed to have done. As Kozulin and others have noted, the procedural aspects of Soviet experiments were generally not reported in great detail within books such as Vygotsky's even though unpublished reports were generally available from the experimenters themselves at the time. Kozulin (1986, p. xxxi) notes that "the studies by Vygotsky's followers have shown that the basic findings are sound, and that argument may arise only as to the interpretation of these findings."

[62] Scribner and Cole's (1981) in-depth research of the effects of literacy among the Vai people of Liberia revealed that formal-schooling, and the kind of literacy developed within that context, seems to be a more likely reason for the production of educated, 'Western' styles of thinking such as abstract syllogistic reasoning. They, among other researchers, have shown that 'traditional' peoples can and do use these kinds of reasoning in familiar and sensible-sounding contexts, and that with sufficient prompting and scaffolding they can produce the 'correct' answers. Barton and Hamilton (1996) suggest that there are multiple pathways

and Luria claimed that these findings provide support for the idea that the psychological tools given by literacy and formal schooling had been used to mediate the way these people used their cognitive skills.

Vygotsky used the term *internalization* to describe this phenomenon of deriving personal and private skills from the self-directed use of social and public skills. Modern Vygotskian researchers argue over whether the notion of *internalization* (the movement of external abilities or strategies to the 'inside') or the concept of *appropriation* (the gradual taking over of elements of jointly carried out skills by the learner from a guide or expert) more accurately captures what goes on with development (see Valsiner, 1997, for a defence of internalization and Rogoff, 1990, pp. 193-197 for advocacy of appropriation). Appropriation fits more comfortably with the distinctly Vygotskian notion of the *zone of proximal* (sometimes *potential*) *development* or *ZPD*. Vygotsky (1930-1935/1978) characterised the ZPD as "the distance between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance or in collaboration with more capable peers." (p. 86). He argued that psychological research should not only pay close attention to the individual abilities of a child but also to those things the child is able to do in interaction with an adult or other guide (this no doubt derived from his interest in education and pedagogy). If a child can accomplish a certain task with the help of someone else, then it lies within the child's zone of proximal development. Vygotsky argued that the child can then be supported to take over an increasing proportion of the task until they can perform it on their own. However if all aspects of the task are beyond the child, that is, lie outside the child's ZPD, then there is no possibility that the child can be guided toward acquiring an independent ability to do the task.

In recent years a number of psychologists have utilised Vygotsky's ideas to understand various aspects of children's cognitive development. One increasingly popular area of research is the exploration of the phenomenon of *private speech* (a.k.a. egocentric speech, self-directed speech). The notion of private speech was central to Vygotsky's deep interest in the relationship between language, as a social symbolic phenomenon, and thought.

---

for developing these abstract cognitive skills, not all of them reliant upon literacy. Rather, different societies can make explicit the meanings in language in different ways. The metalinguistic role of word signs and their relation to meaning are said (by the likes of Olson [1996]) to be made explicit and this explicitness is said to enable many abstract modes of thought. Western society has done this primarily via the expository text, but other societies have different 'ways of speaking' which perform similar expository functions. Notably, none of these critiques deny the basic Vygotskian premise that cultural practices do alter the ways in which we deploy our basic cognitive skills.

Modern Private Speech Research

Vygotsky made many proposals about private speech based upon his own research as well as that of his colleagues. However, his interests were such that many of his ideas were never tested during his short life. Indeed, many of his thoughts about private speech defy any straightforward sort of empirical investigation. Briefly, Vygotsky argued that speech gradually combines with pre-verbal thought, the sort of thought possible, for instance, in chimpanzees, to create verbal thought. According to Vygotsky (1934/1986, chap. 4) verbal thought possesses many of the qualities that are commonly considered to be central to human cognition – qualities such as logical conceptualisation, voluntary memory, and objective representation. Vygotsky (1930-1935/1978) argued that verbal thought derives from a prior ability to use speech in social contexts and that private speech is a transitory stage between overt social speech and silent inner speech.

> [W]hen children find that they are unable to solve a problem by themselves ... [t]hey then turn to an adult, and verbally describe the method that they cannot carry out by themselves. The greatest change in children's capacity to use language as a problem-solving tool takes place somewhat later in their development, when socialized speech (which has been previously used to address an adult) *is turned inward.* Instead of appealing to the adult, children appeal to themselves; language, thus takes an *intrapersonal function* in addition to its *interpersonal use.* When children develop a method of behavior for guiding themselves that had previously been used in relation to another person, when they organize their own activities according to a social form of behavior, they succeed in applying a social attitude to themselves. The history of the process of *the internalization of social speech* is also the history of the socialization of children's practical intellect. (p. 27).

Children experience speech with adults and other behavioural guides (older peers, siblings) in situations of joint activity where the guide verbally instructs and scaffolds the child's actions in their ZPD. In such situations language has the function of channelling and modulating the child's basic cognitive skills. The child gradually takes over this guiding language and uses it to 'tell themselves what to do' (direct attention, inhibit the wrong actions, work out what to do next, and so on). For Vygotskians, speech is both a higher mental function *and* the means by which other elementary skills are organised into higher functions. Talking to oneself forms the basis for bringing basic skills under control.

Recent research has unearthed much supporting evidence for many of Vygotsky's claims, brought into question some hypotheses, and revealed a variety of conflicting findings for other Vygotskian claims (table 3.2 summarises some of these findings). Overall the main thrust of Vygotsky's ideas has received considerable empirical support: private speech *does* seem to develop during a child's preschool years and seems to derive from skill in using social speech; private speech often serves a self-regulatory role, improving performance on effortful, reflective tasks; and over the early school years private speech becomes increasingly internalised – although it appears that it persists into and throughout adulthood in difficult task situations (see, e.g., John-Steiner, 1992).

Importantly for the ideas being mooted here, private language ability seems to play a major role in the kinds of effortful and reflective skills (modelling environments, counterfactual reasoning, thinking about things that are not present) used to negotiate complex cognitive domains. Frauenglass and Diaz (1985) found that more private speech is prompted in children when they attempt to solve semantic tasks (e.g., picture classification, picture ordering) than when they work on perceptual tasks (e.g., puzzles and block design). This suggests that some kind of verbal mediation is required when we are engaged in effortful and reflective tasks as opposed to tasks that can be achieved using experiential cognition (Norman, 1993b) or 'situated seeing'.

Table 3.2: Summary of Recent Private Speech Research Findings
(based on Berk, 1992)

Finding	References[63]
Strong positive correlations between child's level of social interaction and the complexity of their private speech.	Berk & Garvin, 1984; Kohlberg, Yaeger, & Hjertholm, 1968
Development of private speech mediated by verbal and socially responsive home environments.	Berk & Garvin, 1984; Diaz, Neal, & Vachio, 1991
Movement from externalized (out loud) to internalized (inaudible muttering, mouth movements) private speech occurs over preschool and early school years.	Berk, 1986; Berk & Garvin, 1984; Berk & Potts, 1991; Bivens & Berk, 1990; Dickie, 1973; Frauenglass & Diaz, 1985; Klein, 1964; Kohlberg et al., 1968
Level of private speech development correlated highly with general cognitive ability in preschool and early school years (IQ, school ability).	Berner, 1971; Bivens & Berk, 1990; Deutsch & Stein, 1972; Diaz, Padilla, & Weathersby, 1991; Frauenglass & Diaz, 1985; Kleiman, 1974; Kohlberg et al., 1968; Pechman, 1978; Roberts & Tharp, 1980
Amount of private speech displayed increases with difficulty of task.	Beaudichon, 1973; Behrend et al., 1989; Berk & Garvin, 1984; Deutsch & Stein, 1972; Dickie, 1973; Kohlberg et al., 1968; Murray, 1979; Roberts, 1979; Zivin, 1972
Private speech has a self-regulatory function (as well as other possible functions). Level of private speech is correlated with success in various tasks.	Azmitia, 1992; Beaudichon, 1973; Behrend, Rosengren, & Perlmutter, 1992; Bivens & Berk, 1990; Flavell, Beach, & Chinsky, 1966; Goodman, 1981; Keeney, Canizzo, & Flavell, 1967; Klein, 1964; Murray, 1979; Roberts, 1977, 1979
Private speech seems to be a universal feature amongst children.	Berk, 1986; Bivens & Berk, 1990
Private speech is often used to solve difficult (semantic) tasks rather than practical tasks that are solvable using perceptual strategies.	Frauenglass & Diaz, 1985

[63] All references are cited in Berk (1992). The following references were 'in press' in that publication and have been updated in this table: Berk & Potts (1991), Diaz, Neal, and Vachio (1991), and Diaz, Padilla, and Weathersby (1991).

A number of studies support Vygotsky's claim that challenging tasks elicit more private speech than simple tasks or tasks that lie beyond the child's current ability (e.g., Behrend, Rosengren, & Perlmutter, 1989). There has been some debate over the generality of the claim that private speech seems to serve a self-regulatory function when dealing with challenging, off-line kinds of tasks. However, much of the research that challenges this idea derives from experiments that involve only a brief analysis of the children's activity in artificial laboratory situations. Berk (1992) suggests situations such as this may not tap into children's natural and spontaneous problem-solving strategies because they may cause self-consciousness and because there is not the same motivation to tackle problems as there are in more natural settings.

Berk (1986) and Bivens and Berk (1990) found that private speech occurs much more often in natural environments with ability-related challenges when they studied children's mathematics problem-solving behaviours in their classrooms. Indeed in a three year longitudinal study (Bivens & Berk, 1990) they found that individual development in private speech mirrored increases in an ability to sustain attention and concentration while working on academic problems.

Summary of the Sociohistorical Position

Vygotsky's sociohistorical theory suggests that many of our distinctly human cognitive abilities are a joint product of our natural cognitive capacities and the effects of being immersed in social and symbolic human environments. Vygotsky hypothesised that the human psychological development of higher mental functions involves an internalisation of skills and abilities that are initially used in social interaction. One of his core interests was in the way language provides an ability to restructure the way people think. He argued that in private speech children access and utilise the focusing and regulatory functions of social speech in their own individual activity. Adults can guide a child's activity using language and the child can subsequently take over that language and use it as guide in independent contexts. The implication is that much of the content of modern cognitive psychology is not just part of our native neural endowment but in fact is socially and linguistically constituted. Higher mental functions such as explicit memory, problem-solving, and reasoning consist of natural skills that are 'glued together' in the course of social interaction. This means that "[i]f one decomposes a higher mental function into its constituent parts, one finds nothing but the natural, lower skills." (Kozulin, 1986, p. xxv). This theme is at odds with the cognitivist attempt to use formal task analysis to derive cognitive algorithms from analyses of the problems that agents face and to map them onto neurobiological structures and processes.

Summary

In this chapter I have described five different fields that have influenced recent ESD interactionist thought. Although the views of human and animal psychology held by proponents of these different fields are by no means identical, they are united by the fact that they all challenge some or all of the assumptions that are central to cognitivism. Specifically, they all challenge the idea that cognition can best be understood as a purely in-the-head affair that involves the construction of detailed models of the environment and detailed plans for action. The next chapter extracts a collection of shared themes from the insights of the research in these fields. Together the themes provide us with the beginnings of a statement of the core commitments of an interactionist alternative.

4. Embodied, Situated, and Distributed Interactionist Principles

A person is a unique mode of confluences and divergences, a moving locus of interactions.

Susan Oyama (1993, p. 488).

The Main Principles of the Interactionist Approach

The embodied, situated, and distributed interactionist viewpoint that derives from the research traditions described in the last chapter can be understood as a collection of mutually reinforcing principles and assumptions. In this chapter I will describe what I believe are the central shared themes of interactionist researchers.

ESD theorists tend to assume that the most profitable place to start the study of cognition is with as few preconceived boundaries as possible. Thus these researchers tend to focus on (1) the *agent-environment system* as the primary unit of analysis and view all of its components as contributing to the production of cognitive activity as *equal partners*. Because of this emphasis ESD thinkers (2) pay close attention to the ways in which causal influences criss-cross the boundaries of the brain, body, and world. In particular, interactionists commonly analyse cognitive activity in terms of *perception-action cycles*. These assumptions are borne out and reinforced by other findings including (3) the fact that the internal structure of agents (especially their nervous systems) are usefully viewed as *decentralised* (made up of interacting activity layers with no central command and control centre) and essentially *reactive* (they mostly respond directly to environmental 'input' rather than engaging in complex modelling and planning) architectures. In other words, natural-born cognizers excel at hooking directly into their surroundings and producing coordinated action. By contrast people seem to find effortful tasks such as planning, reasoning, and logic more difficult to tackle and slower to achieve. This suggests that the basic design of natural cognition is, as Clark (1997) puts it, to be "bad at logic, good at frisbee". ESD-friendly research thus shows (4) how often complex-looking behaviour is the outcome of this natural 'good at frisbee' architecture coupled to highly structured environments. The difficult bits of cognitive activity (keeping track of past achievements, juxtaposing these achievements, etc.) are often off-loaded on to environmental structures. (5) This coupling, or constant looping of activity, environmental alteration, and perception, results in the *interactive emergence* of complex, dynamic activity. Perhaps the most important implication of interactive emergence is that it shows how complex 'high-level' cognitive activity can arise from the interaction of simple, low-level reflexes (orchestrated by activity layers) and structured surroundings where goals, decisions, and procedures do not exist pre-coded in either internal or external resources. Behaviour, or even behaviour *potential*, can not be 'decoded' just from the structures

inside the agent or just from the contingencies offered by the environment. (6) The behaviour-relevant structure of an agent can only be understood within the context of the environment that the agent lives in and the historical environment that the agent's ancestors inhabited. Similarly the behavioural environment itself must be understood in a relational manner. Because of this many ESD theorists make use of some version of the notion of 'the environment from the point of view of the perceiver' (*effective environments* or *Umwelten*) to describe the 'problems and tasks' that the agent faces. Effective environments describe what the world "*offers* the animal, what it *provides* or *furnishes*, either for good or ill" (Gibson, 1979/1986, p. 127), where such an interpretation derives, not from some sort of subjective interpretation by the agent, but rather from a "creature's morphology, its sensori-motor characteristics, and the activity patterns it performs ..." (Hendriks-Jansen, 1994/1996, p. 285). Thus the description of effective environments helps to show how and why creatures do what they do when in varying conditions.

The rest of this chapter involves a more detailed analysis of these six major interactionist themes.

The Equal Partner Assumption and the Agent-Environment Complex

Central to all interactionist accounts is the idea that cognition goes on inside a system that is substantially larger than an individual mind/brain. This idea involves the ontological claim that "[c]ognitive processes are not located exclusively inside the skin of cognizing organisms" and the epistemological claim that "[i]t is not possible to understand the nature of cognitive processes by focusing exclusively on what is occurring inside the skin of cognizing organisms." (Rowlands, 1999, p. 22). The extended cognitive system is typically thought to be a dynamic one that continously changes size and composition. It consists not only of the agent's brain but also their body, tools and artifacts that they use, the layout of the local physical environment which is often deliberately altered to accommodate particular activities, the local social environment of tutors, colleagues, friends, novices, apprentices, and passersby, and the wider social and cultural environment consisting of social norms, practices, routines, collectively created plans and procedures, discourses, and so on.

The embedded components of these agent-environment systems[64] share an *equal role* in the ongoing production of cognitive activity. No one component is more central or important than another in the production of any particular behaviour. *Epistemic credit* (Clark, 1997, p. 69) should be spread amongst all of the 'components' that play a vital role in the production of cognitive activity no matter which side of the skin they lie on. The

---

[64] Also known as *extended mind systems* (Clark & Chalmers, 1995), *functional systems* (Hutchins, 1995a), and *behavioural systems* (Keijzer, 1997).

idea that we should distribute epistemic credit to bodily and environment, as well as neural, structures amounts to an acknowledgment that the 'information' necessary for the production of activity is also distributed throughout the wider cognitive system. For this reason I refer to these embedded components as *enablers* as they all enable various kinds of cognitive activity. For example, the production of a particular essay within a particular timeframe can be disrupted by both the brain dysfunction and the loss of a valuable reference text. And although some sort of writing may be able to be produced without the text, or indeed with some sort of neurobiological disorder, it would not result in the successful accomplishment of the task at hand. This is not to say that some elements of the typical extended cognitive system are not always at hand or not necessary for all cognitive activity. It is not even to claim that such equal causal influence precludes the existence of principled boundaries between brain, body, and world (Clark, 1998, pp. 513-516). It is simply the claim that that some activities are only possible using particular enablers that reside outside of the body.

Frequent Boundary Crossings

One important implication of moving the focus of inquiry from the mind/brain to the agent-environment system is an increased attention to the subtleties of the dynamic interactions between brain, body, and world. ESD theorists are obliged to pay close attention to the ways in which agent-environment systems hang together as well as to the kinds of systemic patterns that emerge from the coupling of each system's components. In particular, interactionists hold that cognitive processes often include environmental manipulation.

Action Loops and Perception-Action Cycles

The importance of considering the interplay of environment and agent is found in the roughly equivalent ESD notions of the *perception-action cycle* (Goldfield, 1995; Kugler & Turvey, 1987; Reed, 1996; Thelen & Smith, 1994; Wheeler, 1996), *perceptually-guided action* (Varela et al., 1991), *active perception* (Wheeler, 1996) and the *action loop* (Clark, 1997). The basic idea underlying all of these concepts is that cognitive activity, in both its fundamental 'low-level' sense (e.g., moving around, manipulating objects, avoiding obstacles) and its advanced reflective form (e.g., solving problems, constructing artifacts, writing a book), involves constant cycles of information-seeking and environmental manipulation in the service of making new (or altered) information available. Rowlands (1999) suggests that, central to the interactionist perspective, is the *manipulation thesis* which holds that "cognitive processes are, in part, made up of manipulation of relevant structures in the cognizer's environment." (p. 24). Rowlands does not restrict this idea to manual manipulation of objects in the environment but also includes the changes that animals make to the information they have access to when they, for instance, move their

head or eyes. On this reading the movements of sensors in interactive vision, mentioned in the last chapter, constitute a kind of environmental manipulation. Cognition is thus characterised in terms of a continuous interplay of body, brain, and world. Clark (1997, p. 36) puts it nicely when he characterises an action loop as "an intricate and iterated dance in which "pure thought" leads to actions which in turn change or simplify the problems confronting "pure thought"." (p. 36).

This simple idea challenges the cognitivist tendency to view perception and action as the two relatively low profile end points of a cognitive episode. Action (output) and, to a lesser extent, perception (roughly, input) are viewed by the cognitivist as inhabiting the border area between the environment and the 'cognitive processing' of the agent. Moreover, perception and action are typically treated in terms of the *transduction* of environmental energy into 'internal symbolic computation' and vice versa. Cognitivist research overwhelmingly focuses on what goes on between input and output, almost completely ignoring the important ways in which agents continuously explore and manipulate their surroundings. This is the case despite the fact that much of our activity consists of a fine-grained ongoing 'mutual tuning' of body and world that Smith (1996) calls *pure connectedness* and Clark (1997, 1998a) refers to as *continuous reciprocal causation.* Indeed, this sort of deep coupling with the world may form the basis of our more complex temporally-extended activities that are central to most cognitive psychology. For instance, a person solving a classic cognitive psychology puzzle (e.g., the Tower of Hanoi, Cannibals and Missionaries, or the Water Jug problem; see Eysenck & Keane, 1995), playing Scrabble (Kirsh, 1995), or putting together a jigsaw puzzle (Clark, 1997) may be traditionally depicted as working their way through a 'problem space' in one or two complicated off-line ponderings. An ESD-inspired analysis would pay close attention to the often rapid, short duration exchanges of the person exploring and manipulating the puzzle components whilst looking for patterns and points where various actions are appropriate.

One relatively simple task, that demonstrates that environmental manipulation is crucial to successful action, is doing a jigsaw puzzle. Kirsh (1995) uses jigsaw puzzling as an example of the way in which we manipulate our environment to simplify the 'cognitive operations' we must carry out to accomplish a task. When assembling a jigsaw we do not simply stare at all of the pieces lying on the table, build a model of them, and then mentally manipulate them until a plan for putting them together is achieved which can result in the production of the appropriate 'motor program'. Instead we engage in many environmental manipulations such as:

1. Grouping all of the flat edged pieces together so that we can build the edges first.

2. Grouping all of the similar (colour, texture) pieces together (they probably fit into the same area) so that we can make fine-grained comparisons for a smaller area.
3. Physically rotating and fitting pieces into holes, trying them out one by one, rather than trying to figure out which holes they will fit in by eye.

This latter environmental manipulation strategy is central to another cognitive activity – the playing of the computer game Tetris. Kirsh and his colleague Paul Maglio (1994) have carried out a number of experiments using Tetris. The findings again show the importance of environmental manipulation in cognitive processes and not just carrying through actions and the problems with traditional formal task analyses.

The point of Tetris is to prevent various shaped 'zoids', which fall from the top of the screen, from accumulating beyond a certain height. This is achieved by spinning and moving the falling zoids so that they fit together in layers. Once a layer is formed without any gaps it disappears thereby avoiding a game-losing accumulation of zoid piles. Thus Tetris requires the player to work out where a zoid can be best placed to produce gap-less layers.

Kirsh and Maglio, (1994, pp. 518-522) found that Tetris players did not make the kinds of moves that might be expected from a formal task analysis of the 'problems' a player faces in Tetris[65]. In particular, they expected that players would first examine the current layout of the game and the position of the dropping zoid before moving the it. They also suspected that players would only move the zoid in order to put it into position. They could thus calculate the average amount of keypressing needed to achieve this task. However what they found was that players rotated much more often than was necessary for simply placing blocks.

After examining their data and talking to the players Kirsh and Maglio decided that players, especially skilled players, perform two distinct kinds of actions when playing: *pragmatic actions* and *epistemic actions*. Pragmatic actions (or *performatory actions*, [see Reed, 1996]) are deployed to utilise or alter some part of the agent's surroundings in order to achieve some goal or task. For Tetris players dropping a zoid into place is a pragmatic action. Epistemic actions (or *exploratory actions*), by contrast, are aimed at re-presenting some aspect of the world in order to make it easier or possible to perceive a particular

---

[65] In fact Kirsh and Maglio designed a classical SMPA simulation of a Tetris player called *RoboTetris*. RoboTetris worked through a series of four processing phases: Phase one involved creating an early, bitmap representation of selected features of the current situation; phase two involved encoding the bitmap representation in a more compact, chunked, symbolic representation; phase three consisted of computing the best place to put the zoid; and the final phase involved computing the trajectory of moves to achieve the goal placement. (Kirsh & Maglio, 1994, p. 519).

feature or property of the environment. Epistemic actions are performed purely for the purpose of providing new and informative views of manipulated objects so that the uses of the object can be made (more) salient. In the Tetris example players often spun zoids to ascertain whether they would fit into a gap in the accumulating zoid layers and to discover what kind of zoid was available when only a small part of it had appeared from the top of the screen. This was particularly important because several zoids come in normal and mirror-image forms and these forms may not be easily discriminated without physically manipulating them within the time constraints of the game[66]. In addition a number of expert players used 'epistemic translations' as well as rotations for establishing the column position of the zoid.

In addition to this relatively high-level kind of activity other research suggests that this kind of perception-action looping occurs at even more fundamental levels of cognitive activity. Kugler and Turvey (1987; see also Goldfield, 1995) hypothesise that organisms couple with their environments in two ways. The first is with low energy interactions with *flow fields*, such as reflected and ambient light, and the second is with high-energy interactions with *force fields*, such as bodily manipulation of environmental furniture. These two kinds of interaction can be glossed as perception and action respectively. Organisms are 'designed' to couple the former with the latter – which is just a fancy way of saying that natural agents turn low energy stimulation into high-energy interaction in an adaptive manner. A simple example will serve to illustrate the low-level nature of some action loops.

Consider the very basic behaviour of a fly 'deciding' to flap its wings in order to take off (Marr, 1982, pp. 32-33). This, if anything may seem to be a prime candidate for an environmentally-independent process whereby the fly's brain simply instructs the wings to flap in response to some sort of external or internal stimulus. But flies do not use this kind of process. Instead the fly's wing-flapping circuits are wired up to the fly's feet. When the feet are touching a surface the wings remain still but when they cease to be in contact with a surface (typically when the fly 'jumps') a signal triggers the wing-flapping behaviour. McClamrock (1995) describes the external loop in the following way:

> Flight control depends not on signals from the brain to the wings. Instead, the fly exploits the environmental regularity provided by the surface (as well as its own sensory capacities) to run a part of its control loop *outside* its body and through its immediate surroundings. The local surface mediates the signal which, in a slightly roundabout but quite reliable way, makes its way from brain to wings. The fly beats its wings

---

[66] Presumably a player would have to be able to 'mentally rotate' these zoids in order to judge their form if it was not possible to physically rotate them. Kirsh and Maglio (1994) calculated that such mental rotations are much slower to perform than physical epistemic actions. They estimate that it takes 800 to 1200 ms to mentally rotate a zoid 90° but only 100 ms to physically rotate it.

directly in response to a stimulus in the environment - the tactile change at its feet - but also consistently and reliably manipulates that stimulus so as to control flight from the brain as well. (pp. 85-86)

This process is a basic, but robust, action loop *par excellence.*

Interpreting Boundary Crossing

Questions arise as to whether we should interpret these sorts of boundary crossings as 'messages between representational arenas' or as a continuous interplay of resources. The former interpretation can be captured by an extended representationalist strategy that Clark has called *catch-and-toss* (1997) or *interactive* (1996) explanation. The basic idea is that our internal cognitive components should still be understood as representations manipulated according to particular computational rules albeit in a much more attenuated form than those that feature in classical cognitive science. Internal components correspond to *partial* (localised, action-oriented, deictic) representations containing *part* of the content needed for modelling and/or planning cognitive activity. The idea here is that the environment, or possibly a non-neural part of the body, contains some of the structure ('content') that makes an activity possible and, importantly, that this external structure does not need to be copied into the agent's internal structure for it to play such a role. It is rather like the distributed cognition activities where much of the 'information processing' goes on in artifacts outside the body. For instance, a person-and-calculator system can multiply two large numbers, but what is represented inside the agent during such operations are things like 'the first number is 71 986' and 'push the **x** key for multiplication' rather than any detailed mapping of numeric structures – that, in an important sense, goes on inside the calculator. Thus what goes on 'in the head' are essentially 'contentful states' involved in manipulating and perceiving the environment. Of course examples such as this one can be attacked on the basis that it *is* possible to do mental arithmetic in one's head. However others (e.g., Ballard, 1991; Churchland et al., 1994) argue that the 'partial representation' approach applies even to basic low-level activities such as object recognition and related visual processes because these processes also involve frequent boundary crossings. And in situations where boundaries are frequently crossed the activity outcome is grounded in the mechanisms resident on *both* sides of the boundary. Thus 'overall content' (roughly, the representation of the activity situation) is not found only within the head.

Take Marr's fly example mentioned earlier. A proper explanation of the fly's flying behaviour cannot be had just by examining the fly's internal machinery. One could gloss the 'content' of the component that lifts the fly off the surface it rests on as **jump** but its significance in terms of observable and ecologically significant behaviour is that it leads to the fly *flying*. Indeed that is probably why such a component is *in* the fly (courtesy of natural selection) in the first place. So the overall representational content of the activity of flying can be said to reside, not just in the appropriate neural circuits that cause jumping

and wing-flapping-when-my-feet-aren't-touching-anything but also partly in the local environment where the control loop is completed.

The latter non-representational viewpoint is better captured from an *emergentist* and *dynamical coupling* standpoint. Smith (1996) likens this sort of relation between world and agent to the relationship that exists between a jacket and its wearer.

> It would be bizarre to say that the jacket 'detects' your motion--even though, sure enough, it changes states in a way that is lawfully correlated with your motion. It changes state subject to the constraints of overarching physical law because it is *connected* to you--by coming along with you. (Smith, 1996, p. 219)

Under this interpretation of frequent boundary crossing, when an agent and an object in its local environment are 'connected', it is simpler to understand the relationship in terms of coupling than it is to resort to a 'ping-pong' kind of representational explanation where there is a rapid cycling between world changes and partial representational constructions.

The choice as to which interpretation, representational or dynamical, is more correct depends on whether you think that it is still possible to assign at least some sort of minimal content to the physical states of the components within any of the subsystems that make up a behavioural system. Clark (1997) argues that, despite the inroads that ESD-related research has made into the traditional conception of mental representation, it is still useful (and possible) to assign some sort of useful informational or semantic content to the embedded components of a behavioural system:

> Rather than amounting to a clear case against computationalism and representationalism *in general*, what we confront is another body of evidence suggesting that we will not discover the right computational and representational stories unless we give due weight to the role of body and local environment - a role that includes both problem definition and, on occasion, problem solution. ... Nonetheless, this whole complex of important insights is fully compatible with a computational and representational approach to the study of cognition. (p. 154)

However Clark (1997, 1998, pp. 513-516) acknowledges that there exist some forms of boundary crossing where "the spatial envelope demarcating each component [e.g., brain and world] is of little significance if our object is to understand the evolution of the real-world behaviors." (1998, pp. 514-515). He gives as examples a radio receiver and transmitter system that is rigged up so that the receiver modulates the transmitter (1998, p. 514) and a two-'neuron' system that exhibits an oscillatory dynamics even though neither neuron does so by itself (1997, p. 164). In both cases it is impossible to pin down the activity of the overall system to 'instructions' (representations) buried inside just one component. The analogical mapping that he is suggesting here is of the agent onto the receiver or one neuron, and environment onto the transmitter or the other neuron. He speculates that such *continuous reciprocal causation* may well occur in situations such as an improvised jam session by a jazz trio, dancing, interactive sports, and in group conversation. In such cases little explanatory mileage can be made by trying to single out

one component as being the source of input and the other as the 'processor' of that input, because "the target phenomenon is an emergent property of the coupling of the two (perfectly real) components, and should not be "assigned" to either alone." (Clark, 1997, p. 164). He concludes his analysis of agent-environment dynamics with what may seem to be an extraordinary backdown for a confessed representationalist:

> Where the inner and the outer exhibit this kind of continuous, mutually modulatory non-decouplable coevolution, the tools of information-processing decomposition are, I believe, at their weakest. What matters in such cases are the real, temporally rich properties of the ongoing exchange between organism and environment. (Clark, 1997, p. 166)

He argues, however, that taking the representation out of activities involved in 'local environmental contact' does not obviate the need for a representational account of those parts of cognition where the referents of the cognitive activity lie outside the agent's perceptual range. A discussion of these representation hungry situations is of course vital to any developed theory of human cognition and constitutes a major theme of coming chapters. For now, however, it is important to note that there exist robust non-representational candidate explanations for action loop phenomena.

Reactive, Decentralised Architectures

ESD interactionism's focus on the importance of the interactions between brain, body, and world has forced a rethink of hypotheses about the way agents' brains and bodies are put together. By taking the role of the environment more seriously than cognitivism typically does it has become obvious that the place of the agent in the production of activity may also need to change as well. We have seen how situated robotics research is founded upon quite a different architectural philosophy from that of classical AI. And we have seen that situated action theorists cast humans as primarily deploying embodied skills to deal with the contingencies of the immediate and local environment. One of the implications of understanding cognitive processes as partly worldly phenomena is that the traditional formal task description approach of cognitivism will prove to be insufficient for characterising the actual structure and organisation of the 'cognizer's innards' (see my discussion of the emergent functionality problem in chapter 2). Rowlands (1999) calls this implication the *principle of the non-obvious character of evolved internal mechanisms* (NOC). He describes NOC in the following way:

> For the performance of a given task T, and for any internal mechanism M which has evolved in organism O and which, when combined with suitable environmental manipulation on the part of O, allows O to perform T, the nature of M is not always obvious on the basis of T. (p. 81)

Hendriks-Jansen (1996) suggests that NOC is one of the central lessons of situated robotics research and studies in ethology. Within ethology behaviour has often been shown to

derive from a collection of semi-independent *fixed-action patterns* 'triggered' by particular *sign stimuli* in the environment[67]. What glues these activity patterns together into a coherent looking and adaptive coalition is not some internal executive program, but rather the structured layout of the local environment in which the species in question has evolved. Because there tend to exist recurring, well-structured spatial layouts and temporal patterns in these species-typical environments, the animal does not need to 'represent' these relationships in its mind/brain. Indeed, extended patterns of activity made up of environmentally triggered fixed action patterns tend to be *more* robust and adaptive than those that rely upon internal models of the world because they respond directly to the unique layouts of specific environments. No two local environments are ever the same even though they may contain recurring elements (sign stimuli) organised in reasonably regular ways. The similarities between these ethological notions and the ways in which situated robots built with subsumption-architectures work are quite striking. Hendriks-Jansen (1996) argues that ideas from both fields should be brought together to further illuminate the organisation of the innards of natural cognizers.

These ideas all stand in opposition to the classical vision of the agent as a centralised internal modeller and planner. When one distils out the commonalities in these approaches, a list of what can be called *principles of good design* can be formulated. These principles are of use, not only to ESD-inclined AI engineers, but also to those who seek to build models and theoretical frameworks for explaining the activity of natural cognizers. The principles I will discuss here are the *principle of reactivity*, the *007 principle of parsimony,* the *opportunism principle,* and the *soft-assembly and decentralisation principle.*

The Principle of Reactivity: React Rather than Predict

This principle holds that natural cognizers and well-designed robots primarily respond to the contingencies and 'signals' in the immediate, local environment rather than attempting to predict or model distant environments. Situated robotics research has shown that mobots that are designed to pay primary attention to what is happening in the current, local

---

[67] Hendriks-Jansen (1996) describes a fixed action pattern as "a unit of behavior with a rigid motor component and – optionally – an orienting component to direct its course." (p. 219). The classical ethological notion of a fixed action pattern has been modified by more recent expositions in order to downplay the fixed, internal, and innate aspects that Lorenz (e.g., 1957) argued were central. Thus G. Barlow (1968) coined the term *modal action pattern* and Hendriks-Jansen (1996, chap. 12) refers to *emergent activity patterns.* A sign stimulus is an "uncomplicated feature of the environment, such as a particular size, shape, or color, which singly or in combinations, trigger the IRM [innate releasing mechanism]." (p. 219). Lorenz conceived of an innate releasing mechanism as the internal mechanism which responds to the sign stimulus and triggers the activation of the fixed action pattern. All of these concepts have been critiqued within modern ethology (see Hendriks-Jansen, 1996), but the idea that behaviour arises from the environmental activation of a collection of action-pattern mechanisms remains useful.

environment are more robust and adaptable than those that constantly attempt to look ahead. McClamrock (1995) suggests that such an in-built *procrastination* design makes use of the fact that it is often the case that useful information is available from the local environment around the time an agent will need it, often for the simple reason that immediate threats and benefits are of primary importance when agents need to 'decide' to act.

This basic reactive architecture made up of simple reflex-like responses to easily detectable aspects of the mobot's species-typical environment can be subtly modified to anticipate upcoming events by, for instance, building a simple associative conditioning layer on top of the basic reactive architecture (Pfeifer & Verschure, 1992a, 1992b; see also chapter 7). In Pfeifer and Verschure's *Distributed Adaptive Control* architecture this additional conditioning layer (realised using an embodied neural network) leads the robot to anticipate collisions, after a few collision experiences, and take appropriate action before they occur. This sort of modification of a basic reactive architecture differs substantially from the design philosophy of cognitivist-inspired robots in that anticipative behaviour is understood as a kind of diachronically extended reactivity rather than as involving a dedicated forward planning system. Situated subsumption architectures bypass many of the difficulties of internal modelling architectures by avoiding

> the costs of maintaining a rich model of a highly dynamic environment by relying less on predicting and modeling and more on perception. ... We can trade complexity of modeling against speed, reliability, and flexibility in perception. We can take a strategy of waiting longer and looking more; but to do so, you'd better be able to quickly and reliably read the right kinds of properties off the perceptual environment ... In general, we "cheat" off correlations between easily-parsed properties of the environment and the ones that we care about for action. (McClamrock, 1995, pp. 94-95)

Human beings also seem to spend a large amount of their time reacting to local contingencies rather than looking beyond their current situation (see, e.g., Suchman, 1987). Thus humans may not be entirely unlike situated robots in much of their day-to-day activity. McClamrock (1995) gives the following example of human 'reactive cognition'.

> When we drive a car, much of the finer-grained structure of the complex-path we take through traffic - when we stop, when we change lanes, etc. - is dictated by the contingencies of the driving environment. My behavior with respect to any given stop sign is relatively simple; while the complex pattern of stops and starts is a clear reflection of the complex pattern of locations of such signs in the surrounding environment. (p. 84)

What needs to be 'internally planned', if anything, is only a set of rather basic motivating goals such as, in McClamrock's example, "getting across town to the supermarket." The complexity of the finer-grained activity comes from reacting to the local environment. Similarly, Clark (1997) suggests that human cognition is, at base, of the 'good at frisbee, bad at logic' kind – the sort of reactive, perception-action skill that is nicely modelled by intrinsically pattern-recognising and pattern-completing connectionist networks.

The 007 Principle of Parsimony (a.k.a. the Barking Dog Principle)

The second principle arises naturally as a consequence of the first: the well-designed, adaptive and robust agent should off-load as much behavioural control and needed knowledge onto the environment as is possible in order to produce behaviour that is adaptive and efficient (see, e.g., Hendriks-Jansen, 1994/1996). Clark (1989, 1997) refers to this idea as the *007 principle*:

> In general, evolved creatures will neither store nor process information in costly ways when they can use the structure of the environment and their operations upon it as a convenient stand-in for the information-processing operations concerned. That is, know only as much as you need to know to get the job done. (p. 46)

Rowlands (1999) refers to this idea as the *barking dog principle*[68] and develops a detailed evolutionary argument as to why we should expect natural cognizer's to favour environmental manipulation strategies over strategies which demand complex internal mechanisms. He argues that internal mechanisms require a higher investment in energy for both their implementation and maintenance than it does to implement and maintain reactive, environmental manipulation mechanisms. It is thus differentially selectively disadvantageous for an organism to develop an internal mechanism to tackle a task if a manipulation mechanism will do the job just as well. We have already seen this kind of principle at work in situated robotics research and interactive vision research where reactive manipulation strategies do tend to be computationally cheaper compared to classical AI alternatives.

The 007 principle is useful partly because the local environment often contains rich and redundant sources of information that a surrogate internal model may not (see, e.g., Gibson, 1979/1986; McClamrock, 1995, pp. 93-95). The artificial environments created by earlier AI workers[69], that were originally intended to simplify the problems cognitivist-inspired robots face when dealing with unbounded environmental novelty and the exponentially increasing computational demands of planning further and further ahead, can actually make things "too hard because they are too simple." (Chapman, 1990, p. 1 quoted in McClamrock, 1995, p. 93). Norman (1993b) makes the same sort of observation from a more psychological point of view:

---

[68] Clark (1997) calls it the *007 principle* because it suggests that natural cognizers, like spies, should only have access to that knowledge that is necessary for them to carry out their task and no more. Rowlands (1999) derives his *barking dog* principle from interpreting the old adage "why keep a dog if you are going to bark yourself" as "if you've got a dog you won't need to bark yourself."

[69] For example, the computer-modelled block world in which Winograd's 'virtual robot' *Shrdlu* lived and the high-contrast, simply furnished real world environment of *Shakey* (see Copeland, 1993; Dennett, 1991).

With a disembodied intellect, isolated from the world, intelligent behavior requires a tremendous amount of knowledge, lots of deep planning and decision making, and efficient memory storage and retrieval. When the intellect is tightly coupled to the world, decision making and action can take place within the context established by the physical environment, where structures can often act as a distributed intelligence, taking some of the memory and computational burden off the human. (pp. 146-147)

One important finding of situated robotics and interactive vision researchers has been the discovery that real-world environments provide a host of 'cues' (information sources) that mobots can use to latch onto important properties, objects, or events. So, for instance, the coke can collecting mobot *Herbert*, mentioned in the previous chapter (Connell, 1989 cited in Clark, 1997), uses a laser and a video camera to detect a rough outline of a can rather than going through a hugely demanding Marr-like (Marr, 1982) process of computing a context independent 3D object from the light falling on an array of photoreceptors (see also Horswill & Brooks, 1990). Another Brooksian robot (Horswill & Brooks, 1990; Horswill 1992) can detect and track objects using a pair of simple and rather crude visual systems: a 'segmenter' that roughly isolates 'objects' (a large band of spectral frequencies that have no energy in the markings of background surfaces) from the background and a motion detector that manages to (reasonably often) 'stick' to moving objects. Although both of these systems are rather poor at their jobs they work well enough when operated together and in conjunction with the robot's other behaviour layers. Importantly, both the segmenter and motion detector are designed to work in a particular kind of habitat where (roughly) objects stand out from their background (the background texture constraint) and where the robot and other objects rest on the same surface (the ground plane constraint). In other words, the segmenter and the motion detector are designed *on the assumption that* the robot will 'live' in a species-typical environment (i.e., the MIT labs) characterised by these constraints[70].

Moreover, such 'quick and dirty' methods of picking out important objects, events, and properties can be even more efficient when used in *particular contexts*. As McClamrock (1995, pp. 96-97) notes, once a particular object (or whatever) has been picked out in a certain situation an agent no longer needs to have some procedure for picking it out in all possible environments, but only within the restricted array of objects in the current local environment. Thus an agent can use very simple basic cues to 'locate' previously 'identified' objects. And this ability can be further extended by modifying local

[70] This idea that natural cognitive mechanisms are often 'designed' to make efficient use of environmental assumptions is essentially the same kind of claim made by Marr (1982) and, more recently, by evolutionary psychologists (e.g., Barkow et al., 1992). For instance, Ramachandran (1988) argues that the vertebrate eye design assumes that the world is illuminated from above. However, in these reasonably mainstream cognitivist approaches the mechanism itself is not usually considered to be distributed across world and organism as it is in Horswill and Brooks' robots.

environments to intensify simple perceivable differences between items by, for instance, marking items, moving them to and/or keeping them in a standard location, or by attaching them to another noticeable item (essentially an external form of classical conditioning; for a related argument see Dennett, 1993).

This lesson can be fruitfully extended to the behaviour of natural agents by thinking of an animal's species-typical environment as a kind of 'particular context' in which specific perception-action reflexes are specially suited, thereby making much or all perception-action activity largely a use of 'quick and dirty cue usage'. Indeed, Hendriks-Jansen's (1996, pp. 81-86; see also R. Barlow, 1990) discussion of the peculiar visual system[71] of the Horseshoe crab (*Limulus polyphemus*) provides a possible case in point. It turns out that the visual system of *Limulus* exists primarily for the rather specific job of spotting another crab to mate with at night, on a beach, at high tide. And it seems to do this by increasing the receptivity of its ommatidia at night to crab-sized objects moving across its field of vision at the speed at which horseshoe crabs move. The crab does not distinguish 'movement across the visual field' in terms of whether this movement is due to its own motion or that of its target because it does not matter in terms of the 'task' it is faced with. Indeed the entire visual system of the crab seems to capitalise on (and lean rather heavily on) the regularities of one of its species-typical environments.

One implication of thinking in these terms is that it should bring into question the assumption in much cognition research that animals' perception-action systems are primarily in the job of abstracting the same 'objects' from multiple environments, as in Marr's (1982) paradigm. It may well turn out to be the case that many animals, including possibly humans[72] in much of their day to day activity, do not and cannot *see* or *recognise* the 'same' object in two different kinds of environment. Indeed, what would be the

---

[71] 'Peculiar' because, until recently, no one knew just why the crab had a visual system at all. Apparently it does not use it to find food or to detect predators.

[72] Clark (1997, pp. 25-31) and Churchland et al. (1994) provide several examples of how human contact with the environment is skewed toward task-specific ways of 'accessing the world' rather than pulling out a relatively general and value-free objective representation of the external world. Although our phenomenology is often of a clear, objective picture of our surroundings a number of clever experiments have shown that we only access what we need to get by. For instance, by tracking eye movements and tailoring the input to the small part of the visual field focused upon, junk text can be introduced to the other parts of visual field in such a way that it is not even noticed. To the experimental participant it appears that meaningful text exists in the rest of the visual field. The impression of a clear wider picture encompassing our entire visual field, rather than just the 0.01 percent that we foveate, may result from the fact that our "visual perceptual instruments ask and answer their questions so quickly and effortlessly that it seems as though all the answers are already, and contemporaneously, in our minds." (Thomas, 1999, p. 221; see also Churchland et al., 1994, p. 37)

biological point in doing so where the usefulness (or affordance) of an object depends upon the context in which the world is encountered?

So far 'environmental leaning' has been discussed at quite a low level in terms of the use simple mechanisms make of local environmental structure. But as we have already seen people make extensive use of social surroundings to modulate existing cognitive abilities. The notions of *precomputation* (Hutchins, 1995a), *enforcement* (McClamrock, 1995), and *scaffolding*[73] (Clark, 1997) all encompass the idea that people (and perhaps other animals) often deliberately modify their surroundings in order capitalise on their basic (experiential, perception-action, situated, quick-and-dirty) cognitive skills. But not all environmental scaffolding is deliberate in nature. The social activity of agents can also have profound unintended consequences on the behavioural complexity of agents. Hutchins (1995a) illustrates this with a modified version of Simon's (1981) tale of the ant on the beach. Simon (1981) notes that if we follow an ant's movement over a beach it will appear to us that its complicated behaviour is due to some sort of complex 'path planning and implementing' machinery realised in the ant's nervous system. But this would be a mistake, Simon argues, because the complexity of the behaviour viewed can be equally well, and more parsimoniously, explained by taking the ant to have basic sensori-motor reflexes for responding to encounters with twigs, gullies, sandy mounds and so on. In Hutchins' (1995a) parable a *community* of these ants finds itself upon a beach newly smoothed by a recent storm.

> Generations of ants comb the beach. They leave behind them short-lived chemical trails, and where they go they inadvertently move grains of sand as they pass. Over months, paths to likely food sources develop as they are visited again and again by ants following first the short-lived chemical trails of their fellows and later the longer-lived roads produced by a history of heavy ant traffic. (p. 169)

After several months we observe an ant quickly and, apparently intelligently, move about the beach to many food-rich locations. The ant seems smarter than its ancestors of months ago but this cleverness, Hutchins notes, is not due to anything within the ant, but rather to the new 'culturally-shaped' environment of chemical trails and physically scoured out paths. The ant still possesses only a basic reactive architecture, but its behavioural complexity has changed significantly.

---

[73] The notion of *scaffolding*, the process of being guided and tacitly tutored a skill, derives from the work of Jerome Bruner (Wood, Bruner, & Ross, 1976) who was in turn influenced by the likes of Vygotsky and Luria. Bruner originally used the term in the context of the vital supportive environment that caregivers supply children with as they develop language. This language acquisition support system was meant to supplement and contrast with Chomsky's hypothesised innate and internal language acquisition device. Clark (1997) and others have subsequently generalised the notion of scaffolding to refer to *any* kind of environmental feature that plays an important role in the production of cognitive activity.

Hutchins and Hazlehurst have constructed a number of computer simulations that provide some empirical support for the idea that behavioural complexity can be created unintentionally through cultural environmental evolution (Hutchins & Hazlehurst, 1991, 1995; Hazlehurst & Hutchins, 1998). These simulations consist of a group of simple neural nets ('citizens' in a virtual community) that deal with problems in their virtual world either directly, by observing the natural world, or indirectly, via a 'symbolic system' that enables the communication of information about the natural world between agents. The symbolic system begins randomly with each agent deploying a particular sign idiosyncratically to each environmental event. Each agent gets feedback on the success of their efforts to communicate with other agents (obviously two communicating agents who use the same signal for an event will succeed) and eventually a 'socially-agreed upon' signal-event repertoire emerges.

The simulations are run over many generations of agents. Old agents, and the weights on their neural nets, die and new agents are born with the same *initial* network weights as their forebears. Thus, after the initial generation each generation of agents consists of a group of older 'more learned' agents and a group of young ignorant agents. The evolving symbol system (realised in the neural nets of each generation) is passed on from generation to generation from the older agents to the younger ones. This happens despite the possibility of the new agents' ignorant attempts at communicating disturbing the previous generations' gains in communicative consensus.

What Hutchins and Hazlehurst have found is that later generations, despite being structurally identical to earlier generations, are able to solve problems that earlier generations cannot. They are smarter due to changes in the external public resource of the symbol system and not because their innards (basic reactive, pattern-completing architecture) had been improved. Clark (1997) nicely captures the spirit of this research when he writes "advanced cognition depends crucially on our abilities to *dissipate* reasoning ... Our brains make the world smart so we can be dumb in peace." (p. 180).

The Principle of Opportunism

Opportunism, according to McClamrock (1995, p. 87), is "a way to take advantage of the dynamic and unpredictable environment by letting it cue the activation of explicit goals." He gives a homely example of opportunism in terms of a visit to the store: "you take advantage of the opportunity to get milk, because you're at the store, you need milk, and seeing milk at the store reminds you of the goal to get milk." (p. 87). So opportunistic agents can also leave their 'goals' (and 'decisions' and 'choices') to the tender mercies of the environment rather than storing them 'inside' and requiring a complex goal-setting and goal-planning architecture. Or put in more ESD-friendly terms, for the opportunist, goals

arise from the interplay of the layout of the local environment and the bodily states of the agent. Hendriks-Jansen (1994/1996) puts it this way:

> The 'control mechanism' which makes the 'decisions' or 'choices' is thus inextricably embedded in the world. This does not mean that the system as a whole can be conceptualized as a behaviourist stimulus-response engine. Its perceptual and behavioural history will affect its responses to current stimuli, and the same stimulus may elicit different responses on different occasions. In the terminology of dynamical systems theory, perturbation by the environment alters the systems parameters and affects its evolution equation. (p. 285)

Thus, in contrast to the cognitivist agent who is always explicitly trying to satisfy one or more goals (in, perhaps, a roughly Newellian and Simonesque [1972] problem-solving manner) an ESD agent simply 'meshes' its bodily states with what goes on in the surroundings taking advantage of whatever features of the local environment present themselves for use. A simple example may be a mobot that can seek out a power supply socket to recharge itself. It may operate in a Brooksian manner ('wander with intent') until its internal energy supplies drop below a certain threshold level at which point a 'move towards a power socket if sighted' behaviour layer begins to override other behaviour layers in the ongoing 'bidding war' that occurs between the various activity layers that make up the mobot's subsumption architecture. The mobot does not actively seek out a power supply until the environment provides the appropriate cues for initiating the relevant activities. As Hendriks-Jansen (1996) notes in a related discussion of Mataric's (1991) robot Toto:

> If it makes any sense to talk about a decision-making process that activates one reflex rather than another, that process clearly cannot be located inside the robot. To a large extent, it is the particular environment for which the robot was designed that serves as the choice mechanism between the various low-level reflexes.
>
> This does not mean that the robot merely responds to its environment. Its own movement is as much responsible for the behavioral "choices" that result in wall following as is the structure of its environment. If the robot did not move, the environmental contingencies that cause the low-level reflexes to come into play would not arise. (Hendriks-Jansen, 1996, pp. 190-191)[74]

Opportunism is thus one more design principle which can reduce the 'computational burden' of complex tasks such a environmental searching, goal setting, and so on that require a complex spatial and temporal breakdown of the environment to produce internal models of actual and desired environments. Within an appropriately structured environment (e.g., a series of rooms with a number of accessible power sockets) with an appropriately designed agent-architecture (e.g., an early enough low energy warning threshold for the activation of the 'power socket approach' behaviour layer) an agent can 'survive' quite happily without the need for complex, high-level behavioural planning

systems. Of course, environmental change (e.g., a reduction in the amount of available power sockets) could spell curtains for such a situated opportunist, but not perhaps for a planful and thereby more flexible cognitivist agent. However, as I will argue in chapter 8 this more advanced intentional behaviour may be explainable in terms of a social/cultural reuse of a basic, opportunistic, situated architecture without requiring an internal model-building system.

The Principles of Soft Assembly and Decentralisation

Soft assembly is a term used by Thelen and Smith (1994), Goldfield (1995), and others to refer to the activity of a functioning system that constructs itself or self-organises from a collection of coupled subsystems (e.g., nervous system, skeletal or link-segment system, musculotendon system, circulatory system). The activity of each subsystem perturbs other subsystems and ultimately this mutual perturbation leads to the establishment of a stable, usually adaptive, systemic behavioural pattern. Furthermore, when certain parameters are changed within one of the subsystems the overall system will (often) re-organise into a new and different stable state. For instance, the overall behaviour of Brooks' mobots derives from the mutual interaction of the different behaviour layers (subsystems). This is in contrast to 'hard-assembled' systems that derive their ordered behaviour patterns from the constraining effects of a central command and control centre. Similarly, Thelen and Smith (1994) (and others that base their work on the influential ideas of Nicholas Bernstein; see Thelen & Smith, 1994, pp. 75-77) contrast the emerging view of human and animal movement as the outcome of the self-organising, soft assembly of skeletal, neuromuscular, and local environment systems, with the more cognitivist view of movement as the outcome of commands from central pattern generators in the nervous system (see Goldfield, 1995; Saltzman, 1995; Thelen & Smith, 1994).

Maes (1995; see also Clark, 1997) illustrates the difference between a centralised, hard-assembled system and a decentralised, soft-assembled system by contrasting two ways of robustly controlling a multiple-processor computer network so that it efficiently deals with the numerous subtasks that make up its overall job. The classical system has a separate central controller that gathers information from each of the processors and uses various rules and heuristics to work out which jobs should be assigned to each processor. The decentralised system, by contrast, consists only of the multiple machines linked in a clever way. As each machine works it will, from time to time, create new jobs. At such times the machine sends out 'requests' to the other machines for estimates of how long the job will take given its current workload. The 'highest bidder' (the machine that signals the most

---

[74] See Hendriks-Jansen (1996) and Maes (1995) for defences of this view against traditional models of behaviour selection.

efficient delivery of a job) is assigned the job. This soft-assembly system thus self-organises an emergent scheduling system in a computationally-efficient and robust manner; it does not suffer when a machine crashes or is damaged and it is not hampered by model-updating problems that may strike a centralised system.

Soft-assembled systems have the virtue of being able to easily and adaptively deal with a variety of environmental and task conditions (e.g., changes in terrain, damage to muscles or limbs, differing energy demands for moving at different speeds, movement while encumbered by clothes or footwear). In all of these cases a system must carefully 'rebalance' the contribution from the various subsystems in order to continue producing the requisite behaviour. A cat, for instance, can continue to walk, balance and so on with a leg in a cast by subtly altering the kind of gait it uses via different muscular and neural activity. A simple walking robot, by contrast would likely tip over or move in circles unless it could compensate for such changes. Such hard-assembled systems are characterised by brittleness in changing conditions and by the incredible demands on internal 'command and control' complexity when dealing with increasingly dynamic environments. Self-organising systems seem to be able to naturally deal with many kinds of perturbation of the agent's body by either compensating for changes or by moving into a new kind of adaptive, stable state (see also Clark, 1997, pp. 42-45).

Interactive Emergence

In the previous section it was noted that the observed behaviour of environmentally-exploitative systems, such as situated robots and dynamically-coupled systems, is not completely specified in the internal structure of the agent. That is, the structure of the environment plays a very significant role in the shape and structure of the behavioural activity that is produced. This stands in contrast to the view that the agent controls the behavioural outcome via an inner 'kinematic template' (Saltzman, 1995) that specifies the exact form of the behaviour to be produced. Behaviours that cannot be predicted from, or, more importantly, are not solely caused by, the internal structure of the agent's mind/brain are said to be *interactively emergent* (Hendriks-Jansen, 1996; see also Clark, 1997; Steels, 1995)[75].

---

[75] The prefix *interactive* distinguishes this form of emergence from others where a high-level phenomenon occurs by virtue of the interaction of many components *within* a system such as the emergence of orderly convection rolls from the microinteractions of molecules in a pan of heated oil (see Kelso, 1995). This contrast is not meant to imply that within-system emergence does not occur within interactively emergent systems – situated robots with neural network architectures (which are 'within-system emergent') are a case in point.

Patterns of activity whose high-level structure cannot be reduced to specific sequences of movements may emerge from the interactions between simple reflexes and the particular environment to which they are adapted (Hendriks-Jansen, 1996, p. 8). Hendriks-Jansen (1996) notes that in systems whose behaviour is interactively emergent

> [t]here is no differentiation of the activity as a whole into a fixed and a variable component. Indeed, the explanation of why certain activity patterns appear fixed, in the sense of being easily recognizable, does not reduce to a formal representation or generative mechanism for those patterns inside the creature. ... The emergent pattern of activity will not have a correlate in the creature's nervous system, but it can be decomposed into simpler activities, which may themselves have neural correlates, or be further decomposable into even simpler activities. The synthesis of these activities, producing the emergent activity pattern, cannot be paralleled in a corresponding synthesis of their neurophysiological correlates or their mathematical characterizations. Interactive emergence means there exists no overall formal description of the high-level phenomenon, though its pattern will be clearly recognizable within the context of the creature's environment. (Hendriks-Jansen, 1996, pp. 228-229)

Interactive emergence is nicely illustrated by some examples within the ESD-related literature. Hallam and Malcolm (1994 cited in Clark, 1997, pp. 111-112) have suggested a simple design for interactively creating a wall-following behaviour in a robot. Such a robot need only possess an in-built bias to steer slightly to the right and a collision detector (which could be as simple as a springy rod) that signals the robot to turn slightly left for a brief period. Together these two subsystems, in interaction with the environment, could successfully create a wall-following strategy of the 'veer and bounce' variety. A more complex design that uses infrared reflection and deflection to mediate environmental contact has been implemented by Nehmzow, Smithers, and McGonigle (1993 cited in Steels, 1995; see also Smithers, 1992). It essentially makes use of the same two behaviour layers – an object avoidance layer (akin to the collision detector) and a wall-seeking layer (akin to the steering bias). The layers use infrared reflection from a lateral surface (a 'wall') to modulate an 'attraction' (inclination to steer) toward the surface. The closer the mobot gets, the lower the attraction. When tuned adequately the activity of the two layers results in a kind of attractive/repulsive equilibrium that maintains the mobot's wall-following behaviour. This behaviour is emergent "because the category 'equidistance to the (left/right) wall' is not explicitly sensed by the robot or causally used in one of the controlling behavior systems." (Steels, 1995, p. 92). Similar designs are found in the mobots created by Brooks (1991a, 1991b), Mataric (1991), Pfeifer and Verschure (1992a, 1992b). The primary lesson of the ways in which the wall-following behaviour is assembled by these systems is nicely summarised by Clark (1997, pp. 111-112) when he writes "the wall-following behavior ... emerges from the interaction of between the robot and its environment. It is not subserved by any internal state encoding a goal of wall following. We, as external theorists, lay on the wall-following description as a gloss on the overall embedded behavior of the device."

Interactive emergence is also clearly seen in situations where populations of simple reactive and reflexive agents interact with both each other and their surroundings. Craig Reynolds (1987) provides a simple example with his research attempts to model the flocking, herding, and schooling behaviour of various animals by supplying the 'agents' in his computer program (he calls them *Boids*) with three simple 'reflexes': 1) maintain a minimum distance from other objects (including other boids), 2) match velocity with that of neighbouring boids, and 3) move toward the perceived centre of mass of the boid flock. When each boid follows these simple 'rules' complex flocking patterns emerge. These include flying around obstructions in a natural looking way by breaking into subflocks and then reuniting once the obstruction has been negotiated. Notably such behavioural goals are not explicitly 'stored' within any of the boids. Reynolds' boids can flock without having any dedicated 'flocking mechanisms' for splitting, regrouping, subflocking and so on. All of these behavioural patterns are high-level 'side effects' of a low-level agent-environment dynamics (For related research see Stewart [1998, chap. 10] and Hodgins and Brogan [1994]).

A similar simulation by Resnick (1994) attempts to show how termites collect woodchips into piles. Instead of programming each termite to take each chip to a selected spot Resnick imparted each virtual termite with two basic reflexes: if the termite is not carrying anything and it bumps into a chip it picks it up, and if it *is* carrying a chip and bumps into another chip put the chip down. With these two simple rules 2000 chips are sorted into just 34 piles after several thousand steps even though there is no rule specified to avoid picking up chips from existing piles (indeed termites are 'encouraged' to do so) or to build piles. Pile-building occurs because once the last chip is removed from a spot a pile can never be started at the spot again (because a termite must encounter a chip to drop its load). These 'void areas' are agent-environment entities that specify what the termites can and cannot do at those points. Thus pile-building interactively emerges from the layout of the environment and the simple internal structure of each termite.

Kugler and Turvey (1987) examine the behaviour of real termites in a similar way. Some termites build their nest out of elaborate mudball columns and arches. It is unlikely that these termites possess some sort of built-in instructions for creating such structures and it turns out that the columns and arches interactively emerge from the termites basic 'habits' of leaving a chemical trace on each mudball that they deposit and their tendency to drop the balls where they detect the strongest concentration of the trace. The termites start their mudball construction in the same way as Resnick's virtual termites do by creating piles (columns) of mud. However arches emerge when the chemical trace of a nearby column 'overlaps' with that of the column that the termite is on. The result is a skewing of the high point of concentration toward the edge of a termite's column in the direction of the neighbouring column. Gradually, through the efforts of many termite builders, the two

columns will grow together with their upper layers getting closer and closer to each other. The result is an interactively emergent arch. Other nest structures seem to evolve in similar ways.

Luc Steels (1995) has attempted to provide a deeper analysis of these sorts of emergent behaviours in terms of the *controllability* and *visibility* of significant behavioural variables. Steels' first characterisation of emergence involves the distinction between controlled and uncontrolled variables. An uncontrolled variable is simply some aspect of the agent's behaviour that the agent cannot directly influence. Thus an emergent behaviour is a behaviour that is not directly controlled by an internal component. Instead the behaviour emerges from the interactions between the agent's internal components and between the agent and its environment. To give a simple example, one's car's speed is (hopefully) a controlled variable. One (connected appropriately with the car engine) can systematically influence the car's speed. By contrast, the fact that the car rattles ominously at 85 km/h is an uncontrolled variable of the 'automotive system'. It is not possible to directly influence the speed at which the rattling occurs. It does not make sense, for example, to ask the driver to change the 'rattling speed' to 100km/h. Clark (1997, p. 110) gives a similar example of the uncontrolled variable view of emergence that derives from Hofstadter's tale of a computer system that thrashed about when 35 users were hooked into it. Again it makes little sense to ask a computer analyst to shift the 'thrashing number' to, say, 60. 'Thrashing' is an uncontrolled variable – a side-effect of the way an entire system is put together. *Interactive* emergence can be understood in these terms if the parts of the entire system causing the behaviour include aspects of the environment.

Steels also provides an *invisible variable* characterisation of emergence that is meant to provide a basis for the intuition that emergent behaviour is behaviour that *looks* (to a human observer) as if the system (e.g., an agent) is using a particular feature of the environment to modulate its behaviour, but is in fact using some other kind of environmental hook-up. Again an example will help to illustrate this idea. A person-avoiding mobot (say Herbert) may always maintain a 1 metre gap between itself and people wandering about the MIT labs. This may look as if Herbert is somehow explicitly measuring that 1 metre distance (e.g., by extending a tape measure or by calculating distance via sonar returns) and using an explicit internal representation (call it **d**) derived from such sensing to modulate its behaviour (e.g., IF **d**<1 THEN move backwards). In this case **d** is, what Steels calls, a *visible variable*. However, Herbert may in fact only be maintaining that distance indirectly by, for instance, not letting objects take up more than a particular proportion of its visual field (a sort of 'looming detector') or by not letting the sound-level of footfalls exceed a certain level. In both cases **d** is not calculated or used by the robot. It belongs entirely within the observer's cognitive domain. This view of

emergence can be used to describe *interactive* emergence if the mechanisms underlying the assembly of the 'invisible-variable behaviour' include environmental features.

Both of Steels' criteria for deciding whether a behaviour (or other phenomenon) is emergent rest on the basic idea that some patterns of activity are not exclusively determined by a single internal subsystem and cannot be described in terms of the activity of that subsystem. In cases of *interactive* emergence even the entire set of internal subsystems does not exclusively determine the activity that results; rather, it is the way in which the 'agent-side' subsystems interact with the 'environment-side' structures that gives rise to activity.

Two questions should spring to mind given these analyses of the notion of interactive emergence: What are the advantages (and disadvantages) of producing behaviour emergently?, and Why should we believe that the behavioural patterns of animals are interactively emergent? What we find is that answers to the first question partly answer the second question.

The advantages of producing behaviour via interactive emergence (by contrast with central control) have already been covered in the discussion of situated robots and softly-assembled systems. To recap these reactive, decentralised systems produce behaviours that are more *robust, flexible* and *efficient* than traditional centrally-controlled behaviours[76] (see Steels, 1995; Maes, 1995). Task-specific, softly-assembled systems seem to be able to deal well with the problem of creating the same *kind* of behaviour (distal behaviour) in different environments by varying the particular contribution of the agent's proximal behaviour in an appropriate way. Centrally-controlled systems by contrast are likely to provide the same sequence of movements (proximal behaviour) in different environments, thereby creating different kinds of large-scale activity (distal behaviour). So interactive emergence is particularly likely to be operating in the case of activities that are flexibly adapted to the idiosyncrasies of the local environment. It is not too difficult to see that real animal activity is better approximated by the characteristics of interactively emergent, softly-assembled behaviours.

---

[76] Actually Steels (1995) claims that one of the *dis*advantages of emergent behavioural systems is their inefficiency. Unfortunately this remark has no surrounding context, so it is not clear what he means by this. It is possible, from his engineering point of view, that he means it is more difficult to design and build robots that make use emergent behaviours. This would be certainly in keeping with claims of Brooks (1991a) and others in their struggles to adequately tune their mobots so that they work properly. This however is not what I mean by efficiency. I take efficiency to be the efficient use of internal structures within an agent – what others refer to as the reduction of computational burden. As discussed above, architectures that demand long-range, look-ahead and detailed modelling of the environment require complex computation in order to produce adaptive behaviour.

Hendriks-Jansen (1996; see also Rowlands, 1999) has provided an additional evolutionary argument as to why natural cognizers are likely to have reactive and decentralised architectures and thus why it is likely that we need to understand behaviour from an interactive emergence perspective. Natural selection works on organisms-in-their-species-typical-environments (basically on agent-environment systems) which is just a way of saying that *fitness* is defined relative to the challenges an organism faces in its surroundings (competition from other species, environmental adversities, etc.). The bodies (including nervous systems) that will be selected will be those that produce the right kinds of behaviour *within that particular context.* Therefore, the internal components that get selected will form the agent-side portion of the distributed mechanism necessary for producing a particular interactively emergent behaviour[77].

In sum, ESD-related fields seem to constantly turn up examples of behavioural patterns that are better thought of as emerging from low-level agent-environment interactions rather than being 'executed' or 'initiated' by an internal behaviour controlling mechanism. The implications of such findings are profound. Firstly, they suggest that modellers of cognitive activity need to distinguish between uncontrolled variables and controlled variables, invisible variables and visible variables, before launching into the work of describing the 'computational architecture' of a behaving agent. Failure to adequately work out what the agent senses and uses to modulate its behaviour will lead to the wrong internal dynamics being hypothesized (This is the *frame of reference problem* and is discussed more fully in the section on effective environments below).

Secondly, the phenomenon of interactive emergence makes it clear that we cannot even be sure of the *behavioural* function of inner components if we only examine the internal economy of the agent's mind/brain. No adequate characterisation of the inside can be had without an appreciation of its relations to the outside. The next section is devoted to an analysis of this implication of interactive emergence.

---

[77] It is important to distinguish this argument from the related claim, made by evolutionary psychologists (e.g., Barkow et al., 1992), that the mind/brains of natural cognizers will contain cognitive mechanisms designed to deal with the recurring and stable features of their ancestral environments (e.g., Cosmides' [1989] claim that humans have evolved a cheater-detection module). Hendriks-Jansen's (and Rowlands') argument is that, since much behaviour is an emergent outcome of neural, bodily, and environmental constraints, and thus controlled by a distributed mechanism, natural selection will more likely preserve an adaptive *partial* control structure realised in the nervous system and body of the animal rather than a complete, in-the-head control system. The two different claims are not necessarily inconsistent with each other. However, much evolutionary psychological work does give the impression that the hypothesised modules and Darwinian algorithms designed to deal with recurring environmental features involve complete, in-the-head control systems rather than distributed ones.

The Mutual Definition of Agent Architecture, Environment, and Activity

Environment dynamics, agent architecture, and behavioural activity are all vital components of a theory of cognition (see Horswill, 1992; Hendriks-Jansen, 1996). The environment stimulates and 'perturbs' the agent's body, which in turn compensates, producing sentient movement which we call behaviour. This movement results in changes to the environment, and the positioning of the agent in that environment, thereby changing the ways in which the environment affects the agent. These relationships are tight, continuous, ones. Activity, in the sense of movement, is not a separate material entity from the agent's body but describes its changes and realignments within the environmental milieu. It should thus be clear that these notions of agent, environment, and activity cannot be understood to refer to independent explanatory entities. Cognitive theory must understand them as *relational* terms. Yet traditional cognitive science, at base, does treat these explanatory entities as relatively independent. There is an emphasis in cognitivism on the environment as a viewer-independent and objective entity that must be recovered in great detail before meaningful activity can be initiated. This environment is usually understood in one of two ways: as the world described by physics or, more commonly, as the commonsense world of everyday human life (Chemero, 1998a, chap. 5)[78].

Cognitivism also portrays the agent's innards as containing fully-fledged plans of action and models of the world. Behaviour is seen to be an outcome of internal behaviour programs that specify what to do (the fixed component of behaviour). The environment only serves as a set of constraints that may frustrate the workings of this inner behavioural system (the variable component of behaviour). Thus, for the most part behaviour can be understood almost entirely in terms of what is represented within the agent[79].

---

[78] Gibson (1979/1986) and his followers suggest that the environment should be understood at an ecological level of analysis in terms of the behavioural possibilities it affords animals (a common-sense world of all animals). This is a significant improvement over the other two approaches, but it has been criticized for wavering somewhat between providing an objective analysis of environment (relevant to all animals) and advancing a relational, species- or individual-relative notion of the environment (e.g., Costall, 1995; Varela et al., 1991, chap. 9).

[79] Within philosophy of mind these ideas correspond closely to the claims of *internalism,* of which Fodor's (1980) *methodological solipsism* is perhaps the best known variety. Internalism's basic claim is that behaviour can be purely understood in terms of the 'meanings' (or narrow content) stored within the mind/brain. Externalists by contrast (e.g., Lalor, 1997; McClamrock, 1995) argue, after Putnam (1975, p. 227) that "'meanings' just ain't in the *head*!", and that one must view the operation of internal cognitive components as being deeply *context-dependent.* This means that, amongst other things, two identical twins could be in identical brain states, but have different 'mental contents' because of the differing contexts within which they are embedded (the infamous Twin Earth thought experiments are often used to illustrate this point). The concerns of the internalism-externalism debate do not obviously impact on ESD issues

Representations, once they have been constructed within the mind/brain, are understood to be independent of, or more specifically have content independent from, the vagaries of the local environment.

By contrast, ESD interactionists have come to eschew any attempts to think of environment, agent, and activity as anything but mutually defining entities. This is not, for the most part, anything to do with obscure philosophical argumentation, but rather for the practical reasons of understanding how natural agents work and how situated robots should be built. In light of these demands interactionists have come to think of the behaviourally significant environment of the agent as an *Umwelt* (Von Uexküll, 1934/1957) or *effective environment* (Clark, 1997) that is mutually defined by the characteristics of the world and the sensory characteristics of the agent. Similarly, they have begun to think that the roles and functions of the internal components of the agent can only be properly understood in terms of the structure of the environment in which the agent is embedded. Behaviour is understood to emerge interactively and no amount of inspection of the inner economy of mind/brain will deliver up behaviour programs.

Effective Environments

ESD interactionists are making increasing use of the notions of *effective environment* and *Umwelt* in their studies of cognitive activity. Roughly speaking, the terms *effective environment* and *Umwelt* refer to 'the world as it is understood from the perspective of the agent'[80]. An Umwelt is the world that an agent has access to courtesy of the way that its bodily systems, particularly its sensory systems, are wired together. In the words of Merleau-Ponty (1965) "it is the organism itself - according to the proper nature of its receptors, the thresholds of its nerve centers, and the movements of the organs - *which chooses the stimuli in the physical world to which it will be sensitive.*" (p. 13).

Moreover because of the demands of natural selection on sensory systems an *Umwelt* is a description of the world in terms of what the organism can do with it - an *Umwelt* describes

---

because they have usually been directed at the contents of folk psychological propositional attitudes (such as beliefs and desires) (see Keijzer, 1997, pp. 97-106). However Clark and Chalmers (1995) have argued that ESD research should encourage us to adopt an *active externalism* where behaviour is seen to be the product of environmental and bodily structure (and not just a passive context 'interacting' with a folk psychological 'mind'; but see also Lalor, 1997 and McClamrock, 1995 on how no externalism is really all that passive). Clark and Chalmer's active externalism is almost synonymous with the notion of interactive emergence and thus obviously central to ESD ideas. It is not clear however what implications such an approach has for understanding cognition in representational terms.

[80] Similar sorts of ideas are evident in the notions of *ecological level*, *affordance* (Gibson, 1979/1986), *Lebenswelt, frame of reference* (Clancey, 1989; Pfeifer & Verschure, 1992a), *features* (Stewart & Cohen, 1997), and *enaction* (Varela et al., 1991).

a *space of possibilities for action.* Thus instead of seeing an 'objective' collection of objects, properties, and situations (e.g., a red block 10 metres to the north) animals, including humans, see the world in terms of *affordances* (e.g., a thing to hide behind before that predator sees me). Animals do not see a valueless object and then, once a value-free representation has been constructed, mentally implant it with meaning and value (see, e.g., Smythe, 1992a). Rather, the structure of an organism's sensory systems (in the broadest sense) permits interaction with the world-as-it-is in a particular way that is determined by the survival and reproduction requirements of that organism. The objective world is sampled and aspects of it are accessed in ways that are useful to the specific animal[81]. A couple of examples may help to make these ideas clearer.

Perhaps the most well-rehearsed example comes from Von Uexküll's (1934/1957) analysis of the behaviour of the tick – an animal that seeks out warm mammalian skin for laying its eggs. According to Von Uexküll the tick's entire perceptual and cognitive world (its *Umwelt*) consists of just three distinct entities (objects, properties, features) that correspond to what we know as butyric acid (found on mammalian skin and thus a sign of nearby food), tactile contact with mammalian skin, and body heat. Each stimulus sets off a particular behaviour: butyric acid causes the tick to drop onto its potential host, tactile contact is the sign to move about in search of heat, and heat is the cue to burrow into the skin in search of a place to reproduce.

The effective environment of the coral octopus has been described as consisting of "'food' (small things that move or are carried in forceps) and 'non-food' (everything else). However, the 'everything else' category is subdivided into 'useful crevices' (big enough to squeeze into, too small for a predator to follow), 'home crevice' (I live here), and 'the rest of geography' (not distinguished)." (Stewart & Cohen, 1997, p. 166). Stewart and Cohen call these divisions of an animal's effective environment 'features' (pretty much equivalent to the Gibsonian notion of an affordance[82]).

---

[81] This is not to say that we humans, for instance, cannot and do not further interpret what we perceive in particular ways according to our 'high-level' plans and expectations (for instance, interpreting someone's lateness as indicative of their disorganised character). The main point being made here is that even our most basic, subpersonal encounters are, in a sense, interpretations of the world-as-it-is. There is no fundamental objective contact with this mind-independent reality.

[82] Gibson (1979/1986) thought of an affordance as a agent-relative feature of an object or event in the environment. Objects and events may have multiple affordances (a chair may afford sitting on and standing on to get something off the top of a shelf for a person) and these affordances may not always be obvious or noticeable (the term 'attensity' has been used by Gibsonians to capture the idea that an affordance may be more or less noticeable - an ergonomically designed tool should have a higher attensity than a badly designed one). Thus Gibson was convinced that affordances are objective and real entities (not just 'things in animals' minds') that can be detected because evolution sees to it that an animal is in contact with important

Finally Hendriks-Jansen (1996, p. 254) examines the early interactions of an infant with its caretaker (usually its mother) from an Umwelt perspective. Human infants, it turns out, are blessed with a variety of abilities that, in combination with their need to be closely monitored due to their relative helplessness, create an Umwelt centred on their mothers' faces and the myriad events that occur there. These 'abilities' include poor visual depth of field and early visual fixation patterns. Hendriks-Jansen's example raises a couple of important points. Firstly, that an agent's Umwelt gradually changes with development, and secondly that early, immature Umwelten may play an important focusing and scaffolding role for the development of later skills and distinctions. Infants, so Hendriks-Jansen argues, are designed so that their early behavioural capacities single out important sources of stimulation (e.g., mother's vocalisations, mother's gaze) for the development of later abilities (e.g., development of dialogue and turntaking) (Hendriks-Jansen, 1996, chap. 15).

Thus Umwelten reflect a perspective on the world skewed toward what an animal needs from the world in order to behave adaptively in its species-typical environment. From an adult human perspective the Umwelt of other animal species may seem to 1) exclude or include aspects of the world, 2) group dissimilar things or distinguish similar things, or more subtly, 3) accentuate or diminish regions within perceptual continuities.

A simple example of inclusion/exclusion can be seen in the fact that many insect species are perceptually sensitive to ultraviolet radiation, whereas we are not. Varela et al. (1991, pp. 180-184) provide a more complex example; they point out that while human vision is trichromatic, being facilitated by three types of photoreceptors (for hue, saturation, and brightness), some other animal species have fewer or more dimensions to their 'colour' spaces. Rabbits, squirrels, tree shrews, and some New World monkeys, for instance, have dichromatic vision, whereas goldfish, pigeons, and ducks are tetrachromats. All animals 'see' the same world-as-it-is, but they all have profoundly different perceptual worlds of 'colour'. Notice how the insect-vision example differs from the colour-space one. In the former insects seem to have access to part of the electromagnetic spectrum that we do not. Whereas in the latter the very same electromagnetic radiation has different perceptual consequences depending on the anatomy of the respective animal's visual systems.

Dennett (1991) nicely summarises the idea that animals' perceptual worlds may seem to split apart or glue together, what may seem to us to be peculiar combinations of objects and

---

(for fitness) parts of the world. He argues that a set of affordances should not be understood to be a phenomenal environment: "This can be taken erroneously to be the "private world" in which the species is supposed to live, the "subjective world," or the world of "consciousness." The behavior of observers depends on their perception of their environment, surely enough, but this does not mean that their behavior depends on a so-called private or subjective or conscious environment. The organism depends on its environment for its life, but the environment does not depend on the organism for its existence." (Gibson, 1979/86, p. 129)

events, when he writes: "If some creature's life depended on lumping together the moon, blue cheese, and bicycles, you can be pretty sure that Mother Nature would find a way for it to 'see' these as 'intuitively just the same kind of thing.'" (p. 381). A slightly less colourful, but realistic example comes from the Umwelt of the mantis shrimp (*Squilla mantis*). At a particular stage of its life cycle (the *antizoea*) the mantis shrimp treats *anything* less than half its size which moves as food (Stewart & Cohen, 1997, pp. 165-167). To us floating bits of seaweed, plastic, wood, as well as small fish, do not seem very similar objects. Ironically many of the things that the shrimp sees as food (and thus attacks in an attempt to snip off any appendages) are not nutritious at all.

More subtly, other species' Umwelten may *accentuate* or *magnify* regions within, what appears in other schemes to be, various perceptual continuities, and thereby diminish the 'size' of other differences. That is, what appears to us, or scientific measurement, to be changing in a linear fashion may be importantly non-linear to another species. And other animals' linearity may appear non-linear to us. Our ability to perceive phonemes in a fairly discontinuous manner (when the actual acoustic features, such as voice-onset time or onset frequency of the lowest formant, change linearly) provides one example (see, e.g., Eimas, 1985; Wood, 1976). One could also imagine that heightened sensitivity to parts of the electromagnetic spectrum may provide access to subtle changes in colour that signal different kinds of food quality. By the standards of other schemes such non-linearities serve to strongly tie animals to important parts of the environment so that adaptive behaviour can be produced.

The combination of these features shows how Umwelten 'constitute the natural joints of the world' for animals. An Umwelt is a specific selection of 'bits' of the world-in-itself and also a useful transform of those 'bits' accessed. And this 'categorising scheme' of the world is most likely to make rough sense in terms of the behaviours that a species needs to produce in order to survive and reproduce (in the ancestral environments in which the species' body structures were selected). No doubt there is much room for 'noise' and 'inaccuracy' even within such schemes. For instance, a particular animal may see 'a dangerous situation that requires fleeing' in many situations that do not require such behaviour without it adversely affecting the animal's survival and chances of reproduction. As long as a behaviour that is based upon a particular perceptual scheme does not *jeopardise* survival and reproduction, the mechanisms underlying it can be selected for. Natural selection, as Varela et al. (1991, chap. 9) forcefully point out, deals in viability not optimality.

But even though each agent's perceptual systems arrange the world in different ways according to its needs and capabilities, the Umwelt concept does not imply a complete rejection of any kind of objectivity. An animal that possesses a particular Umwelt can still act in an 'objective' (rational, adaptive, realistic) manner by accurately tracking changes

signalled *within* that scheme as the world about it changes. Since most significant (i.e., important for survival and reproduction) changes in the real world will be detected as changes within the Umwelt, it will provide the animal with the appropriate information required to act adaptively.

The Frame of Reference Problem

Interest in the effective environment concept has arisen because ESD researchers have become increasingly aware that patterns of activity can potentially be produced by a variety of underlying mechanisms. That is, there is a difficulty in working out whether an objectively-described ('viewpoint-independent' or 'third-person perspective') aspect of the environment that an agent *seems* to use to modulate its behaviour is a *visible* or *invisible* variable (Steels, 1995).

As we have seen already, the work of situated roboticists provides a number of nice cases in point. A variety of intelligent-looking behaviours have been simulated in robots using subsumption architectures rather than the more traditional sorts of architectures that are based on a formalised folk-psychological analysis of behaviour. The nagging worry that has arisen from these findings is that perhaps cognitive science's traditional formal task – analytic approach to constructing 'cognitive architectures' (whether they be in theories about human cognition or in plans for building intelligent agents) is fundamentally flawed and unlikely to enlighten us about the ways cognitive beings are put together. Functionalism's grand claim that we need only analyse the multiply-realisable, algorithmic level of cognitive architecture comes into question. In order to deal with these problems ESD researchers have tried to make more explicit, what Clancey (1989; see also Pfeifer & Verschure, 1992a) has called, the *frame of reference problem.* This is the problem of working out what it is that an agent is tuned to in their surroundings (what constitutes its effective environment) and how it uses those sensitivities to modulate its behaviour. Or, in Steels' (1995) terminology, it is the problem of discovering the controlled variables in a behavioural system.

Pfeifer and Verschure (1992a) illustrate the sorts of difficulties associated with the frame of reference problem with an example from work of Franceschini, Pichon, and Blanes (1991) that attempts to model the behaviour of flies.

> [T]he fly only has motion detectors, i.e., if the fly is not moving and nothing moves in the environment, the fly is 'blind'. Let us assume that the fly has a mechanism for turning if it is approaching something fast. Now we are inclined to think that we need an adaptive mechanism for adjusting the distance at which the fly should turn to avoid the obstacle: if it is flying fast it should turn earlier than if it is flying more slowly. However, since the fly is equipped with motion detectors the case where it is approaching a nearby object slowly is - from the point of view of the fly - identical to one where the fly is flying fast and the object is further away. Indeed, the action to be taken is exactly the same in both cases. It should be noted that only for us, the observers, are there two cases. (p. 320)

Thus, Pfeifer and Verschure show that it is vital to understand what constitutes an animal's world (its *Umwelt*) before plunging into hypotheses about the mechanisms that underlie its behavioural patterns. Cognitivism, it is widely argued by ESD interactionists, is none too careful about such issues[83]. There is a growing suspicion among interactionists that a proper understanding of the *Umwelten* of animals will reveal a number of holes in the representational-computational position.

Understanding the Role of Internal Mechanisms

Interactionists not only regularly make the theoretically significant notion of the environment relative to the characteristics of the agent, they also commonly think of the role (function, content, 'meaning') of the agent's inner components as an intrinsically context-dependent affair. As was noted in the previous section dealing with frequent boundary crossings, ESD *representationalists* like Clark (1997), propose that it makes sense to talk of inner components as containing a kind of partial content. These partial internal representations have been called *action-oriented representations* (or AORs, Clark, 1997), *pushmi-pullyu representations* (Millikan, 1996 cited in Clark, 1997), and *deictic representations* (Agre & Chapman, 1987; see Chemero, 1998a, pp. 154-157 for a discussion of all of these). Clark (1997) describes AORs as "representations that simultaneously describe aspects of the world and prescribe possible actions, and are poised between pure control structures and passive representations of external reality" (Clark, 1997, p. 49) and that they "are not neutral descriptions of the world so much as activity-bound specifications of potential modes of action and intervention." (p. 51). In other words AORs are simply mechanisms that make use of the cues and minimal world slices captured by a creature's specially sensitised (to survival – relevant features) sensory systems. The internal 'computations' do not work on sense-data or representations of wordly features (e.g., 'a red box' or 'the slowly moving object to the north') but rather upon the 'interpretations' the organism's sensory systems have made of the world (e.g., 'the opportunity for energy recharging within six steps' or 'the-thing-I-just-kept-away-from').The primary rationale for postulating entities such as AORs is to argue that an agent's inner components relate to (and 'capture' in their content) personally relevant aspects of the environment required to directly modulate activity.

However, Clark's enthusiasm for continuing to think of agent's inner components as representational in nature is not shared by all ESD theorists. ESD *interactionists* (such as

---

[83] Although a strong advocate of the information-processing approach to cognition, Marr (1982) makes a very similar point when he argues that psychological research should begin by paying attention to the computational understanding of a psychological phenomenon (see my discussion of formal task description in chapter 2).

Hendriks-Jansen, 1996 and Keijzer, 1997) argue that it makes little sense to try and think of internal components as representing the outside world. Continuous, mutual modulation between internal components and environmental structure cannot be adequately captured using the idea of representational content (see the arguments advanced in the section on frequent boundary crossings earlier in this chapter). This is partly the case because the operation and function (roughly, content) of inner components of agents are *context-dependent* (McClamrock, 1995). What a component actually contributes to an activity depends upon the state of other bodily components and the state of the environment. Clark (1996a) himself is aware that homuncular (componential) analysis will not provide us with much illumination when we study non-homogenous systems where "the contributions of the parts are highly inter-defined, that is, the role of a component C at time t_1 is determined by (and helps determine) the roles of the other components at t_1, and may even contribute quite differently at a time t_2, courtesy of the complex feedback and feedforward links to other sub-systems." (p. 269).

What we strike when we begin to accept that the 'content' of specific inner component states depends upon where the component is situated within the overall brain-body-environment system is the distinct lack of utility of the content notion. Developmental biologists have already struck these problems when trying to maintain the notion of 'genetic content' (i.e., what kind of phenotypic traits are specified by stretches of DNA) in the face of evidence that phenotypes with identical genomes can vary depending upon variations in the cytoplasmic environment as well as differences in habitat in which the organism develops[84]. Theodosius Dobzhansky introduced the notion of the *norm of reaction* to try to deal with this problem (see Gray, 1992, pp. 172-174). The basic idea is to map out the phenotype as a joint product of the environment and the genome. This may serve as a reasonable approach for organisms, such as plants, that tend to develop in a single habitat but, as Gray (1992) and Ho (1986) point out, animals (and even many plants) alter their surroundings and also often move into different habitats. Because of this, organismic activity, as well as the layout of the environment, must enter into a complete causal explanation of the construction of a phenotype. Even in cases where a particular phenotypic trait is strongly canalised (i.e., reliably appears in a large range of environments) the equation of genetic structure with the primary informational content for

---

[84] Within a gene-focused framework this finding can be understood in terms of the switching on of different genetic programs. Notice, however, that this implies some sort of sensitivity to the environment by the organism's genotype. So the environment again enters into the phenotype construction equation. For a thorough thrashing out of gene-centred versus developmental systems views see (in the chronological order of the debate) Sterelny & Kitcher (1988), Gray (1992), Griffiths & Gray (1994), Sterelny, Smith, & Dickison (1996), and Griffiths & Gray (1997).

building the phenotype is a fundamental misconception. Griffiths and Gray (1994) note that such a belief in an 'informational asymmetry' between genes and environmental resources is an unjustifiable prejudice.

Thus modern developmental biology has begun to portray development as a dynamic, multiple-resourced process (e.g., Gottlieb, 1992, 1998; Griffiths & Gray, 1994; Oyama, 1985). The idea that 'informational content' for the shape of the 'final' phenotype pre-exists within any particular developmental resource (i.e., *implicit preformationism*) is rejected in favour of the notion that "information has its own ontogeny". This is Susan Oyama's (1985) claim that the way a resource or structure 'informs' the outcome of the entire system is a continuously changing quality dependent upon the mutual modulatory effects of other relevant resources (i.e., *probabilistic epigenesis,* Gottlieb, 1992, 1998). She captures this idea nicely with the following analogy:

> An old rhyme goes, "For want of a nail the shoe was lost, for want of a shoe the horse was lost, for want of the horse the rider was lost. ..." Loss of a particular rider could even lead to losing an entire battle. To know whether that nail makes the difference between losing and winning the battle, shouldn't we know what kind of battle it is, on what terrain it is being joined, what the command structure is, and who the opponent is? *Even if it could be shown that the loss of a battle were traceable to a lost nail, this would not make the nail an adequate causal explanation for the entire complex of events that constituted the battle. Indeed, it is the entire complex that defines the nail's role.* (Oyama, 1989, p. 28, italics added)

Such claims do not obviate the need to understand how genetic structures contribute to phenotypic outcomes. Rather, they focus our attention on the necessity of understanding the complex dynamics of an entire developing system as well as making it clear that representational thinking (assigning phenotypic content to particular genetic structures) is at best a coarse approximation of the actual workings of the phenotype-building system. Griffiths and Gray (1997) argue, after Johnston (1987), that "[t]he idea that developmental information resides in the genes is a shorthand for the idea that if all other elements of the developmental matrix are held constant, changes in the genes are reflected in changes in the phenotype." (p. 471). However, they go on to note that "it is equally true that if everything including the genes is held constant, changes in other elements of the matrix are reflected in changes in the phenotype." (p. 472). Thus, it would be wrong to privilege 'genetic information' over any other kind of developmental information in a theory of phenotypic construction. Developmental systems theorists such as Griffiths, Gray, Oyama, Johnston, and Gottlieb argue that we should understand the development of the phenotype as an outcome of the self-organising properties of a system made up of multiple resources such as "[g]enes, methylation patterns, membrane templates, cytoplasmic gradients, centrioles, nests, parental care, habitats and cultures ..." (Griffiths & Gray, 1997, p. 471).

The point of this biological interlude is to argue, as Keijzer (1997, chap. 5, 1998a) does, that this very same pattern of thinking is relevant to our understanding of the role of

agent's inner components in the production of behavioural activity. Interactionists argue that, just as it is a gross explanatory oversimplification to map genes onto phenotypic traits, so it is with inner components (e.g., neural assemblies) and behavioural patterns. Keijzer (1997, 1998a) argues that inner components of animals that promote long-range anticipative behaviour should be understood as *internal control parameters* (ICPs) in the same way that modern developmental biology understands the role of genes in the construction of phenotypes. ICPs are not representations in the classic sense because macroscopic order (behaviour) is not somehow pre-encoded in them. Rather, ICPs play a central role in regulatory networks that give rise to ordered patterns of behaviour over a variety of time scales. In dynamical systems terminology, a control parameter is an aspect of a system that can be altered and that kicks the system into different states or patterns of activity as its magnitude is varied (see chapter 5 for more details). Keijzer proposes that this is just what ICP neural circuits do; they are local (within the body) regulatory networks that make small-scale changes to larger self-organising systems (e.g., the behaving body, social groups, environmental systems of various kinds) so that large-scale, and often temporally extended, order occurs in these larger systems. In other words, neural circuits provide 'simple' magnitude changes to parameters of larger scale systems such that their behaviour changes. Natural selection has seen to it that such magnitude changes reliably produce 'adaptive order' in these large-scale systems (provided they are of a similar makeup to those systems that existed when the small-scale regulatory networks were originally selected). From this dynamical point of view, then, the entire explanatory point of assigning (behavioural) content to inner components vanishes[85]. Inner components 'trigger' rather than 'instruct' behaviour. And this triggering capacity is itself modulated by the dynamics of other systems. For the interactionist enthusiast work must proceed along a different path to that of their cognitivist colleague.

Mutual Definitions

There are two important methodological consequences of embracing the previous collection of interactionist themes. The first is that we must systematically explore agents' behaviours in order to discover what it is that makes up their worlds (Chemero, 1998, p. 181). We cannot assume that the furniture of our commonsense ontology or even the ontology of physics provide the 'input' for cognition. Instead we must contend with, what Clancey (1989) calls, the *frame of reference problem*, the problem of working out what an

---

[85] Keijzer notes that although it may not be impossible to redefine representations in a manner consistent with the nature of ICPs, this further distances the notion of representation from its original explanatory role as an inner isomorphism (of a state of the world or a behavioural plan). It is better, he argues, to admit that we are now talking about quite different entities, that work in quite different ways, and that must be studied using new empirical methods. A new vocabulary will help to signal these quite fundamental changes.

agent is sensitive to. This approach "must combine ethology, neuroscience, evolutionary biology, with elements of more traditional cognitive science to discover the nature of the *Umwelten* of non-human animals and artificially evolved robots." (Chemero, 1998, p.181).

The second important implication of thinking in terms of mutual definitions is that we must be very careful how we think about the functions of agent's inner components in the overall behavioural system. To think of these components as part of a representational or information-processing system is to buy into the idea that, at a subpersonal and physical level, agents take in something from, or of, the environment and manipulate it in order to produce behaviour. But if we are true to a naturalist and materialist philosophy we need to realise that it is no easy task to assign informational content to some privileged parts of the agent-environment system but not to others. We are better served, say interactionists, by thinking that *all* relevant structures help to 'inform' behaviour production and to avoid freezing pre-specified content into any part of the behaviour generating system.

Explaining Wild Cognition

An ESD approach that takes to heart the previously described themes requires a collection of explanatory tools that can do justice to the vision of cognition as consisting of a complex interplay of internal and external resources. As we have seen, traditional computational representationalism takes the task of cognitive science to be the investigation of the role of internal mental complexity in making sense of meaningless inputs and using them to produce plans or instructions for rational behaviour. Information theory and its modern extensions (e.g., Dretske, 1988; Markman & Dietrich, 1998) and computational theory thus form a suitable backdrop for tracking the flow and transformation of information through a physical system. On this view the cognitive project is entirely concerned with what goes on between the sensory surfaces and the motor surfaces; behaviour, environmental activity and structure are of (distant) secondary importance.

The concerns of interactionism go beyond those of cognitivism and are not easily assimilated into an information-processing or computational-representational framework (see, e.g., Keijzer, 1997, pp. 126-141[86]). It is not impossible to view some external resources (e.g., books, pictures, recordings) in representational and computational terms (as e.g., Hutchins, 1995a and Perkins, 1993 do), at least in a loose or metaphorical sense, but such attempts at extending the cognitivist framework face several significant problems: the

[86] Keijzer (1997, p. 126) summarises this idea "To put it bluntly, [cognitivism's] core idea is that behavior is the result of internal instructions, while the interactionist core idea is that behavior is *not* the result of internal instructions."

problem of cognitivism's intrinsic internalism; the problem of turning all external resources into some form of representation; and the problem of defining representation.

Cognitivism's Intrinsic Internalism

The cognitivist explanatory framework is an intrinsically internalist one. It focuses upon the role of the cognizer's innards in producing cognitive activity. The basic premise of cognitivism is that behaviour is caused via an internal interpretation of the outside world rather than through a basic kind of coupling with that world. Any cognitivist attempt to think of external resources as equal partners in the production of cognitive activity must face the charge that it ignores the fact that those same external resources must be re-presented on the inside in order to impact on behaviour production. Put simply, cognitivism is essentially an inside-focused framework and, by adding on a few 'situated enhancements', it renders problematic its central premise. Cognitivism posits explanations of both low-level and high-level cognitive activity by declaring that "a set of representations has to be manipulated until they provide a set of instructions that specify what a behaving system should do to accomplish a goal." (Keijzer, 1997, p. 127). Interactionism does not make such assumptions. Consequently Keijzer (1997) is pessimistic about the prospects of merging interactionist insights with cognitivism.

> [Cognitivism] states that the structure present in behavior results from previously stored instructions. The embodied approach states that the structure present in behavior results from the ever ongoing interaction between organism and environment. It is hard to see how the latter can be given a place within [cognitivism], or how they could be otherwise combined. The only plausible way for [cognitivism] to accommodate interactionism is by turning itself into a different conceptual framework. This move strikes me as a sensible thing to do, but the result will not be [cognitivism] any more. We should keep our terminology crisp here and not call this a way to salvage the existing conceptual framework. (pp. 127-128)

'Representationalizing' External Resources

The main thrust of extended cognitivist analyses such as those of Perkins (1993) and, in some respects, Hutchins (1995a, 1995b) is to continue to think of the cognitive system as a representational-computational system that is, however, distributed over elements that lie beyond the skull and the skin. However, such analyses tend only to easily incorporate *some* significant external resources within a modified cognitivist framework. Other parts of the environment, that interactionists appeal to as genuine parts of an agent-environment behaviour generation system, are not so easily assimilated. For instance, it is easy to think of the elements of the external world as part of an extended computational system when the objects of research are mainly instances of everyday computation such as in Hutchins' studies of position fixing in military navigation (Hutchins, 1995a) and the calculation of airspeed in commercial airliners (Hutchins, 1995b). In these cases the external resources *really are* representations (of the everyday variety) of various parameters and variables central to what are basically arithmetic tasks. Such strategies become less useful when

considering less 'computational' kinds of activity such as doing a jigsaw puzzle (Kirsch, 1995), learning how to ski by using a manual (Costall & Leudar, 1996), or learning how to eat correctly in a middle-class American household (Valsiner, 1997). In these cases it stretches the representational metaphor to think of, for example 'walls that robots bang in to' or 'the layout of jigsaw pieces on a table', as items that contain or transform information about aspects of the world. These things *are* the world (or parts of it) and thus do not need representing at all. They do not contain information even though agents can be informed by attending to them. Clearly then, the computational approach does not easily incorporate all of those environmental resources that interactionists are willing to consider part of behavioural systems.

Defining Representation

Finally, it must be remembered that cognitivists generally believe that the power of external representation derives from the fact that people and other animals possess evolutionarily more basic internal representations. Thus, any adequate theory that explains the cognitive significance of external representation depends upon an adequate theory of internal representation (this is another consequence of cognitivism's intrinsic internalism). However, as noted in earlier chapters, there currently exists no widely agreed upon or even roughly complete theory of content for internal or mental representations. Thus, the status of any *extended* cognitivist approach rests upon the internal representational project being successfully completed. We have the right to be sceptical of 'extended cognitivism' until that time.

An Interactionist Alternative?

ESD-related research and the assumptions and claims shared by the different fields that contribute to ESD thinking provide us with a hodge-podge of thought provoking challenges to many of the common ideas in cognitivism. It remains to be seen, however, whether these challenges will result in a fundamental paradigm shift within cognitive science or will merely form the basis of a more sophisticated cognitivism. It is notable that a number of the claims made above have been resisted by traditionalists, usually by arguing that ESD advocates have overstated their case. Examples of such rejoinders include the following:

1. We may be able to get by without representations in many cases but we still need them for much high-level cognitive activity (e.g., Clark & Toribio, 1994; Kirsh, 1991).
2. There are possible, even likely, explanations of activity that do not require the use of either traditional symbolic representations or connectionist subsymbolic representations, but these alternatives nonetheless imply some kind of representation (e.g., Bechtel, 1998; Chemero, 1998; Clark, 1997; Markman & Dietrich, 1998).

3. A good deal of cognitive activity requires external resources but this does not mean that internal resources are not of primary importance and are not, in a fundamental sense, independent of the environment (e.g., Fodor, 1987).

4. Decentralised and reactive architectures may be smarter than we had thought but there is no evidence that they can be scaled up so that they might account for complex intelligence (e.g., Kirsh, 1991).

At times in this chapter I have addressed some of these criticisms directly, but I have done little more than suggest that some ESD ideas seem more attractive than the cognitivist alternatives. A stronger case for interactionism must rest upon an argument as to why cognitive systems are better viewed from an interactionist perspective. If the representational-computational and information-processing accounts are unsuited to explaining the expanded view of cognitive activity, an interactionist approach is going to need some new set of explanatory tools to deal with issues such as equal partnership and the continuous, modulatory nature of brain-body-environment dynamics, the interactive emergence of activity patterns from the low-level activity of decentralised and reactive architectures situated in complex, real-world environments, and the relational nature of both internal component function ('meaning'), and the stimulating environment or *Umwelt.*

If natural cognizers really do not fit the cognitivist vision of informavores, that is, of systems that pull out and use environmental information in an inner symbol manipulation theatre, what sort of vision of 'living organisation' should we adopt? If a brain is not computing over representations of environmental objects, then what *is* it doing? ESD-interactionists, as opposed to their ESD-representationalist cousins, are currently in the process of formulating explanatory tools for answering these sorts of questions. At the moment the field, such as it is, is fairly turbulent with many different suggestions being pursued by different researchers. Generally speaking, most interactionists would agree with Clark's claim that they are all advocates of, what he calls, the *Thesis of Radical Embodied Cognition.* This is the view that "[s]tructured, symbolic, representational, and computational views of cognition are mistaken. Embodied cognition is best studied by means of noncomputational and nonrepresentational ideas and explanatory schemes involving, e.g., the tools of Dynamical Systems theory." (Clark, 1997, p. 148). In the next three chapters I will map out what I believe to be a promising approach for putting some flesh on the bones of this basic idea.

5. Explaining Wild Cognition with Dynamical Systems Theory

[T]he mathematical theory of dynamical systems is no more a theory of autonomous agents than is the formal theory of computation. Rather, like computation, dynamical systems theory is best seen as offering a conceptual framework for thinking about complex systems, and a framework that is very different from that offered by computation. Where a computational language suggests that complex but highly structured behavior arises from the step-by-step transformation of discrete symbols by identifiable functional modules, a dynamical language suggests that such behavior can arise as a global property of the continuous interaction of many distributed, cooperative processes. Our task as scientists is to use the language and mathematical tools offered by dynamical systems theory to develop theories of particular phenomena of interest.

Randall Beer (1995b, p. 130).

Introduction

The previous chapter gathered together a cluster of phenomena and ideas that interactionist researchers encounter and use in their studies of cognitive activity. It is clear that the research agenda of these scientists differs substantially from the usual fare of mainstream cognitive psychologists and their cognitive science kin. In this chapter I want to suggest that this cluster of themes and findings demand a different set of tools and a different explanatory framework from the computational-representational approach of cognitivism; simply extending the current cognitivist orthodoxy will not suffice. The alternative framework has to somehow make better sense of ESD themes than cognitivism does.

I want to suggest that the tools necessary for studying cognition from an interactionist perspective are already at hand, although they are, as yet, at an early stage of development. In this chapter I will argue that an embodied, situated, and distributed interactionist alternative will likely be grounded in an explanatory framework that uses dynamical systems theory to model cognizers as *autonomous systems* (discussed in chapter 6). This framework provides an understanding of cognitive activity in terms of the coupling of an organism and its environment, an understanding of the organism as a self-organising and self-producing system, and an understanding of nervous systems as regulatory and coordinative control systems rather than as internal world-modelling systems. Such a dynamical systems methodological and explanatory framework, although still far from rigorous and complete, provides a strong and cohesive scaffold for making further sense of the ESD themes mentioned in the previous chapter.

The Promise of Dynamical Systems Theory

Dynamical Systems Theory and Cognitive Science

A number of claims have been made about the ways in which dynamical systems theory can be fruitfully incorporated into cognitive science. The most general claim is that dynamical systems theory is useful for studying cognition because it is the explanatory framework of choice for understanding physical systems that change over time. And since cognitive agents are physical systems that change over time, they should be amenable to a dynamical system interpretation. Because dynamical systems theory is the most general framework applicable to the study of cognition, it carries with it the least number of theoretical preconceptions about the structure and dynamics of cognitive systems, unlike the more restrictive representational-computational approach. Therefore the dynamical approach should be preferred for being more 'open-minded' about the way cognitive systems are built (Beer, 1995a, 1995b; Husbands et al., 1995). Importantly, those who make this point also take the dynamical systems approach to cognition to be quite general in nature and to encompass both digital computational and connectionist systems within the broad class of dynamical systems.

Van Gelder (1995, 1998) argues that this equation of dynamical systems with all state-determined systems that evolve over time obscures the distinctive claims that a non-cognitivist dynamical approach might make about the cognition. Thus, less generally, van Gelder (1998) advances the *dynamical hypothesis* – the claim that, at the *highest relevant level of analysis*, agents are dynamical systems and not digital computers (as is claimed by the *computational hypothesis* central to cognitivism) and thus dynamical analyses and not digital computational ones are appropriate for cognitive science. By this van Gelder seems to mean that the characteristics of cognizers cannot be accurately captured as algorithmic transformations of discrete symbolic states even at the highest, most abstract level of description. This stands in contrast to the claims of Fodor, Newell, and Dennett who claim that such physical symbol systems may be merely implemented by dynamical or connectionist architectures[87]. According to van Gelder and Port (1995) cognition is better captured by a dynamical approach because of the basic facts

> that cognitive processes always unfold in real time; that their behaviors are pervaded by *both* continuities and discreteness; that they are composed of multiple subsystems which are simultaneously active and interacting; that their distinctive kinds of structure and complexity are not present from the very first moment, but emerge over time; that cognitive processes operate over many time scales, and events at different time scales interact; and that they are embedded in a real body and environment. The dynamical approach provides a natural framework for the description and explanation of phenomena with these broad properties. The

[87] It is important to note that Fodor (1975, 1987), Dennett (1978, 1991, 1996) and Newell (1990) each hold quite different views about how such an explanatory feat is possible.

computational approach, by contrast, either ignores them entirely or handles them only in clumsy, ad hoc ways. (van Gelder & Port, 1995, p. 18)

It is important to realise, however, that Van Gelder's dynamical hypothesis suggests a much less circumscribed approach to studying and understanding cognition than does the framework of ESD interactionism. Indeed, dynamical notions have been embraced by theorists who continue to endorse the use of representations, albeit in modified form, and continue to focus upon mostly in-the-head issues (e.g., Elman, 1995) as well as those who lean in a more interactionist direction. However, this is not to say that dynamical concepts do not support an interactionist approach to cognition or facilitate interpreting cognitive activity in an interactionist manner. On the contrary, there are several aspects of dynamical analyses which provide powerful ways of reinterpreting cognitive phenomena that are consistent with the themes mentioned in the previous chapter.

Thus, Keijzer (1997) and others (e.g., Beer, 1995a, 1995b, 2000) focus on dynamical systems theory's facility for explaining the interaction of subsystems (bodily, neural, environmental) using the non-information-theoretic notion of *dynamical coupling*. On this reading dynamical systems theory should be preferred because it avoids the difficulties, such as the symbol grounding problem, associated with a 'message passing and message transforming' approach to the understanding of cognitive interactions. Keijzer (1997) ultimately challenges this reason for adopting dynamical systems theory on the grounds that the notion of dynamical coupling does not and cannot provide a unique understanding of what makes a system cognitive. Instead he argues, along with a number of other writers (Beer, 1995b; Hendriks-Jansen, 1996; Keijzer et al., 1998), that dynamical systems theory itself cannot provide an adequate alternative framework for making sense of cognitive phenomena. Rather dynamical systems theory provides tools that seem, at this stage, to *support* and *encourage* an interactionist approach, rather than providing a complete alternative explanatory framework in itself. The principles of ESD interactionism, rather than those dynamical systems theory, provide the theoretical meat for a challenger to the hegemony of cognitivism.

Dynamical Systems Theory: A Numerical Modelling Strategy

At its most basic, DST can be used to explain cognitive phenomena by thinking of agents as systems that are describable as collections of variables that can take on numerical values that change value over time according to some rule of evolution – an equation known as a *dynamic*. More specifically, one can follow van Gelder (1998) and think of a real world entity (e.g., a person, a tree or a planet) as a *concrete object*. A *system*, by contrast is simply a set of variables that change interdependently. A concrete object can *instantiate* a number of *concrete systems*. For instance, a person can be understood as a reasoning system when we study this particular phenomenon or a kinetic system when we study how bodies fall. The mathematical descriptions (or models) of concrete systems are *abstract*

systems. Van Gelder talks of concrete systems *realising* abstract systems. Thus dynamical analyses of particular phenomena involve the construction of abstract dynamical system models.

A dynamical systems model begins by conceiving of the behaviour of a system at a point in time as a set of values for a collection of *state variables* and any relevant *parameters*. Although variables and parameters appear to play a similar role in these geometric descriptions, they differ in an important way. The changing value of a state variable is dependent upon the values of other variables and parameters. However, while parameters can influence the values of variables they can not be influenced by them and thus parameters usually have a fixed value (gravity is a good example of a parameter that influences many physical systems). Thus variables are interdependent and parameters are not (see van Gelder, 1998)[88].

This collection of variable and parameter values constitutes a *state* of the system. Thus, the entire behaviour of the system at any one point in time can be represented as a single point in a *n*-dimensional hyperspace where each dimension corresponds to one of the state variables. The mathematical rule (typically expressed in terms of a differential or difference equation[89]) that determines the way that the state of a system evolves over time[90] is known as the *dynamic*. As the behaviour of the system changes over time a path (called a *solution trajectory*) is traced out in the space. A space that represents all of the possible states of the system is known as the system's *phase space*[91]. A phase space diagram that contains the set of all (or many) of the possible trajectories that the system can traverse (the

---

[88] Because a system is technically a set of *variables* changing interdependently, parameters are not actually part of a dynamical system (van Gelder, 1998).

[89] Differential equations are used to find the instantaneous rate of change of a variable (the derivative) at a particular value (e.g., we may use a differential equation to calculate the speed of an object at a particular time given information about the object's changing distance over a period of time) (see, e.g., Capra, 1998, pp. 114-118 for a simple exposition; Norton, 1995 for a more mathematical introduction). Importantly differential equations are not always solvable (especially in complex and irregularly behaving systems). Indeed it was this lack of solvability that led to Poincaré developing the ideas that later became central to modern dynamical systems theory (Capra, 1998, pp. 119-126; Norton, 1995, pp. 45-47). When differential equations cannot be solved difference equations are often used instead. These are simply equations that, when iterated many times, trace out the behavioural trajectory of a system. Such equations effectively provide rules for working out the value of x at $t+1$ based on knowledge of the value of x at t (for more detail on the difference between differential and difference equations see Norton, 1995).

[90] The time set used may be discrete or continuous (Beer, 2000, p. 21).

[91] Writers often use the terms *phase space* and *state space* interchangeably. van Gelder and Port (1995, p. 7-9) argue that the phase space is the numerical and mathematical counterpart of the system's real world *state space* or *state set*.

flow), when starting from different initial conditions, is called a *phase portrait.* A diagram that shows these trajectories in terms of the instantaneous direction and magnitude of change at each point[92] is known as a *vector field.*

Describing the Features of Phase Space

A system's phase portrait represents that possible dynamics (behavioural trajectories) of the system and often exhibits interesting behavioural patterns that can be mathematically and qualitatively described. Of particular interest are regions of phase space toward which many trajectories converge. These are known as *attractors.* The space of all initial points which lead to an attractor is known as a *basin of attraction.* Attractors come in different varieties: a *point (or fixed) attractor* is simply a point at which a group of trajectories converge; *a cyclic (or periodic) attractor* is a regular series of points which trajectories move toward and then traverse in an ongoing cycle of activity; and a *strange (or chaotic) attractor* is a series of points that trajectories traverse that exhibit a non-repeating but rule-governed (i.e., non-random) kind of pattern (see, e.g., Elman et al., 1996, pp. 211-218 for some simple mathematical examples). The *stability* of an attractor (that is, the likelihood that a system will seek out and stay in the behavioural regime related to the attractor) is indicated by the size of the attractor's basin of attraction and the attractor's *local relaxation time.* The local relaxation time corresponds to the time it takes for a system to return to a stable state when the system is perturbed by its own intrinsic noise or an outside influence (see Thelen and Smith, 1994, pp. 65-66). In contrast to attractors, areas of phase space that trajectories avoid and flow away from are known as *repellors.* Areas of phase space that attract some trajectories and repel others are known as *saddles.*

Changes to the System Dynamic

The field of synergetics (an off-shoot of dynamical systems theory) pioneered by the laser physicist Hermann Haken (e.g., Haken, 1987; see Capra, 1998, pp. 89-92; Thelen & Smith, 1994, pp. 54-56 for brief overviews) makes use of two kinds of parameter for discussing the ways in which systems made up of a number of component entities come to exhibit this kind of global order (examples include, laser light being made up of the light waves emitted by many atoms, the ordered rolling of heated water made up of a multitude of water molecules, the flow of a crowd through a football stadium).

An *order parameter* (or *collective variable*) describes the modes of behaviour that the collection of components exhibits[93]. For instance, in the finger waggling experiment

---

[92] Differential equations provide this kind of information. Thus vector fields are the natural geometric plot describing these equations.

discussed in the previous chapter, Kelso and his colleagues (see Kelso, 1995) tracked the behaviour of the finger waggling system in terms of a single order parameter that signified the relative phase of the two fingers rather than using a number of variables that describe the independent motions and angles of the two fingers and their joints.

A *control parameter* (this really is a parameter in the sense of an 'outside influence' on the system) describes a particular aspect of the world that can be altered so that it systematically affects the system's phase portrait. Recall that a phase portrait depicts a system's behavioural landscape showing how any initial state will evolve over time. Attractors (and other kinds of limit sets such as repellors and saddles) correspond to regular kinds of behaviour that a system is often drawn toward or pushed away from. However, even this behavioural landscape of attractors and repellors can change when the system is perturbed by forces that lie outside of the system. As noted earlier, these outside forces, that are constant within a particular phase portrait, are known as *parameters*. When a parameter changes it produces changes in the behavioural landscape. Often such changes are only subtle, linear, and continuous changes to the trajectories and attractors (and other limit sets) in the phase portrait. The portrait remains qualitatively similar (or *topologically equivalent* or *homeomorphic*) to the old one (Beer, 1995a). Such subtle change is known as *parametric change* (Thelen & Smith, 1994, p. 63). However, within nonlinear systems[94] a small continuous change in a parameter can lead to a dramatic discontinuous, qualitative change in a system's dynamics (behavioural landscape). Such *phase shifts* (or *phase*

---

[93] It may appear that an order parameter is not a parameter at all as it seems to describe emergent patterns rather than an outside influence of the system. The use of the synonym *collective variable* seems to further muddy the waters. However there are rather complex reasons for thinking of an order parameter as a parameter or 'outside influence' (see Kelso, 1995 for an introduction). Roughly speaking, an order parameter represents the collective 'force' of coordinated parts of the system. "An order parameter is created by the coordination between the parts, but in turn influences the behavior of the parts." (Kelso, 1995, p. 16). Clark (1997) calls this effect *continuous reciprocal causation* and illustrates it in terms of crowd behaviour where "the actions of individuals in a crowd combine to initiate a rush in one direction, and ... that activity then sucks and molds the activity of undecided individuals and maintains and reinforces the direction of collective motion." (pp. 107-108). This kind of phenomenon occurs in all nonequilibrium systems made up of many interacting components such as lasers, heated liquids, and living things. The next chapter examines the concept of a self-organising system in more detail.

[94] Capra (1998, pp. 121-122) and Stewart (1997) note that nonlinearity is extremely common in nature and that the sense that this may not be the case derives from the fact that many nonlinear phenomena have been 'linearised' in the physical sciences in the past in an attempt to make them more tractable. Most nonlinear differential equations are analytically unsolvable (Port & van Gelder, 1995, p. 575) and this difficulty is one the primary reasons why Poincaré and others embarked upon the qualitative study of nonlinear systems (van Gelder & Port, 1995, p. 14); that is, the "study of long-term general features of a dynamical system without attempting to derive or predict specific numerical values of the variables." (Port & van Gelder, 1995, p. 576)

transitions or *bifurcations*) can radically alter the dynamics of a system by 1) the creation or annihilation of an attractor or other limit set (a *catastrophic bifurcation*), 2) the transformation of one kind of attractor into another (a *subtle bifurcation*), or 3) suddenly changing the magnitude of an attractor (an *explosive bifurcation*) (see Abraham, Abraham, & Shaw, 1991, pp. II-86 - II-97). Thus, within nonlinear dynamical systems, small changes in a parameter can lead to discontinuous change. This is the sort of change that is often taken to be an indication that a new causal variable (e.g., a matured neural system or a new environmental agent) has begun to influence the system. Dynamical systems theory provides the tools to understand how a change in a parameter can cause a system to radically re-organise as opposed to being 'instructed' to behave in a certain way by a new causal agent.

The term *control parameter* is typically used to describe a parameter that can be 'twiddled' by an outside agent to control the behavioural landscape of the system. Technically any parameter can serve as a control parameter. However, only a very few control parameters may lead to a behaviourally significant way of controlling the system's behaviour. Such a sufficiently useful control parameter could be used to push the system through a series of appropriate (useful or adaptive) behavioural regimes. For instance, when an animal is faced with a threat from a small sized opponent (reflected in terms of a low control parameter value) one kind of behavioural approach (e.g., aggressive display) is appropriate (signified by the phase portrait given for that value of the parameter). However, when the size of the opponent (control parameter value) passes a certain critical point, a new behavioural regime (e.g., a phase portrait that describes fleeing behaviour) appears.

According to Abraham et al. (1991) a system that can control a control parameter that impacts upon itself through some kind of feedback loop is known as a *self-organising system*. Note that technically any aspect of a system that is modulated by the system itself is a state variable not a parameter. Thus, strictly speaking, a system cannot systematically control one of its own control parameters[95]. However there does exist an interesting sense in which a control parameter can be understood as being reciprocally connected to a system. Dynamical systems theorists understand this phenomenon in terms of the *coupling* of two dynamical systems.

Interacting systems

Two or more systems are *coupled* when the state variable of one system plays the role of a parameter in the other and vice versa. This mathematical formalism nicely captures the

---

[95] In synergetics only systems that are made up of a large number of non-linearly interacting components, and are far-from thermal equilibrium, can be self-organising systems (e.g., Kelso, 1995). I present these ideas in more detail in the next chapter.

idea that two or more systems may modulate each other's behaviours. Because dynamical systems theory theorises the activity of a system in terms of the evolution of continuous numerical values, coupling differs from the traditional cognitivist notion of 'discrete cycles of input and output'. Instead of understanding the relation of two systems in a ping-pong manner, dynamicists speak of *mutual direct dependence* (van Gelder, 1998). This captures the idea that there is an *ongoing* and *instantaneous* modulatory relationship between two systems rather than a series of 'message passing' events between two arenas.

The notion of coupling also illustrates, what van Gelder (1998) calls, the semi-arbitrariness of systems. This arises because any group of coupled systems (e.g., systems *A* and *E*) can also be understood as single larger system (e.g., system *U*) (Beer, 1995b; Wheeler, 1996). In a larger system such as *U* parameters in the coupled systems such as *A* and *E* are effectively converted into state variables.

Dynamical Analyses and Psychological Phenomena

DST describes and explains cognitive systems in *numerical terms*; the variables can correspond to any kind of entity thought to be relevant by the theorist, be they neuronal activation levels, bodily positions and angles, attitude levels, or mood states. The numerical value of a variable at any point in time thus corresponds to the strength, position, or quality of an entity. Such a notion should not be alien to most psychologists who have traditionally tried to provide some systematic numerical analysis of psychological phenomena using chronometrics, scale ratings, and various other methods of measurement.

However, often the variables used in dynamical analyses are of a rather abstract nature representing higher-order invariants, emergent properties, relations between components and so on. Saltzman (1995), for instance, uses abstract constriction types as variables in his dynamical analyses of speech production rather than more physical variables such as lip aperture and lip protrusion. This is partly because DST equations quickly become difficult to visualise and analyse when they include more than a few variables, although, as Beer (2000) notes, it is not impossible and this problem can be circumvented by the judicious analysis of highly influential variables. But the use of abstract variables also occurs because they often provide the most economical and accurate description of multi-component systems with relatively few degrees of freedom. Thus, the quantitative interpretations of psychological phenomena psychologists typically make may not be well-suited to dynamical analyses[96].

---

[96] An additional problem with the quantitative analysis of most modern psychology is that they are simply do not address psychological phenomena dynamically. Often psychological measurements are atemporal or only very vaguely temporal, such as pre- and post-treatment measures.

DST is thus an explanatory framework with a minimal commitment to any particular theoretical framework for understanding psychological phenomena compared to symbolic and even connectionist accounts. Symbolic accounts, as we have seen, consider cognitively meaningful units (e.g., concepts) to be encoded in a unique physical structure which is manipulated and transformed purely according to its physical shape and structure. Encoding is digital and change is discrete. While connectionist accounts admit a degree of fuzziness with regards to the structure-content relationship, and view change in terms of the systematic varying of numerical values, they commit the theorist to understanding cognitive activity in terms of interactions among multiple homogenous components (see van Gelder, 1995, pp. 365-374 for an analysis of the differences between the three frameworks). DST is much less prescriptive than either of these frameworks. This feature of DST accounts has been heralded both as its primary strength and its major weakness (often, ironically, by the same author). Beer (1995a, 2000) argues that DST is a less restrictive framework than that of symbolism or connectionism and should be preferred for just this reason. We are not obliged to make too many a priori theoretical commitments using a framework that can admit any kind or number of variables (even if, in practice, this may prove difficult or intractable) as long as: 1) they have numerical values (whether discrete or continuous), 2) they change over time (whether discretely or continuously), and 3) they always possess a value (even if it is zero).

At the same time a whole raft of theorists (e.g., Bechtel, 1998; Clark, 1997; Eliasmith, 1996; Hendriks-Jansen, 1996; Keijzer, 1997) argue that it is precisely this lack of theoretical commitment that rules out DST as a stand-alone explanatory theory of cognition. DST, argue these critics, provides a covering-law explanation rather than a causal-mechanistic explanation of phenomena (Bechtel, 1998, see below). Without this latter kind of explanatory enterprise we cannot build or model cognitive systems (see Clark, 1997) or utilise useful explanatory rules of thumb such as the decomposition and localisation heuristics (Bechtel, 1998). I will discuss these arguments later on in this chapter. For now, however, I want to focus on the ways in which dynamical systems notions reinforce the principles central to the kind of ESD interactionist thinking canvassed in the previous chapter. In particular DST ideas seem to provide efficient tools for making sense of the ESD claims that:

1. Non-neural (bodily and environmental) resources should figure prominently in descriptions of cognition;
2. Boundary crossings between these different arenas cannot be adequately understood within a traditional information-processing framework;
3. The distinct 'shape' of behaviour emerges from the coupling of neural, bodily, and environmental systems, and cannot be usefully understood as being primarily

controlled by one particular arena (see discussion of Damasio's work in the next chapter).

Attractive Features of Dynamical Analyses for ESD Interactionism

The dynamical systems framework does not provide a ready-made palliative to all of the ailments of current cognitive theorising. Indeed the application of dynamical systems theory ideas to the extremely complicated problems of cognitive activity may require the use of new and more advanced dynamical tools than currently exist (see, e.g., Ho, 1993; Jaeger, 1998; Swenson & Turvey, 1991). But even at the qualitative and metaphorical levels (see Van Gelder & Port, 1995) dynamical systems theory, if used judiciously, can provide some profound insights into ways of understanding the complex phenomena associated with cognition.

It is not difficult to see that dynamical systems ideas dovetail quite nicely with the principles that have arisen from our analysis of ESD research. In this section I will briefly point out what I believe are the significant points of contact.

A Common Calculus for Agent-Environment System Components

The dynamical systems framework provides a common calculus for describing systems both within and outside of the agent (van Gelder, 1998; Clark, 1997). Environmental features, bodily structures, and neural architectures can all be understood in terms of numerical variables that are related to other variables by rules of evolution (i.e., the system dynamic). This stands in contrast to the cognitivist framework which cannot easily extend its representational and computational vocabulary to non-mind/brain systems (see discussion at the end of the previous chapter). This feature of the dynamical view of cognition makes it easy to think of agent-environment systems as important units of analysis and of brain, body, and world as equal partners in the production of cognitive activity. Although a broad dynamical approach to cognition does not require theorists to take into account the *embeddedness* of the brain in bodily and environmental systems, Beer (1995b) argues (after Ashby, 1960) that we should use the following coupled equations representing coupled, nonautonomous, continuous-time, dynamical systems as our basic template for describing cognitive activity.

$$\dot{x}_A = A(x_A; S(x_E))$$

$$\dot{x}_E = E(x_E; M(x_A))$$

In these equations $\dot{x}_A$ is the change in state of the agent and $\dot{x}_E$ is the change in state of the environment. A is the function of the agent's influence on 'input' to the agent system and E performs the same role for the environment system. The variables x_A and x_E are the current state of the agent and the environment respectively. $S(x_E)$ represents the effects of

the environment on the agent via sensory systems and $M(x_A)$ the effects of the agent on the environment via motor systems. Beer notes that environment terms can correspond to the actual environment or the body depending on what aspect of activity is being examined. If one follows Keijzer's (1997) basic interactionist schema *A* could be understood as a system that is itself made up of two coupled dynamical subsystems representing the nervous system (*N*) and the body (*B*).

It should be noted that, even though DST makes it easy to represent agent and environment properties using the same calculus, it does not supply the theorist with any guidance as to the particular metrics that should be used to measure aspects of agent and environment dynamics. In a slightly different context Flach (2000) refers to this task as the *common currency problem*. He argues that mainstream cognitive psychology does not provide us with a common framework for describing the furniture of the environment, the tasks and goals facing the agent, and the control systems that intervene between them. For instance, because of the disciplinary differences between physics, aeronautical engineering, and aviation psychology an understanding of a pilot's task of turning a plane so that it is comfortable for passengers would typically be described in terms of gravity, acceleration, air pressure, wind speed, kinetic energy, work and so on (the environment as described by physics), flap angle, pitch, roll, elevation (in terms of the planes control systems as described in engineering terms), and in terms of mental representations of goals and values of the pilot (the province of psychology). Although all of these terms could be described in a numerical manner (and thus modelled in terms of coupled dynamical systems) no systematic agent-tool-environment dynamics may become apparent. Instead Flach advocates measuring the dynamics of all of the subsystems in terms of an objective, agent-environment vocabulary. The aerial environment could be measured in terms of its contribution to smooth flying. The control systems of the aircraft could also represent banking in terms of 'objective comfort levels' of the passengers as could the task goals. This can be done relatively easily by building measurement systems and control systems that register *higher-order invariants* or *emergent properties*. For instance, a 'comfortable turn' may be measurable using some composite property that depends on the aircraft's acceleration, banking angle, and the characteristics of passengers sitting in cabin seats. This higher-order property would then reflect the amount of and suddenness of swaying felt by passengers. Not only would such a measurement system provide a more easily graspable and analysable description of the agent-environment system, it would also provide a natural set of parameters for designing relevant and ergonomic control systems. If a pilot can directly read off and control the 'comfort level' of a manoeuvre from the cockpit instruments the amount of in-the-head calculating and on-the-fly experimenting will be reduced.

Boundary Crossings as Coupling, not Message-Passing

DST provides a framework for understanding the 'boundary crossings' between body and brain, agent and environment, and indeed different neural systems, that avoids the use of the often problematic metaphors of communication and information transfer and the associated concepts of *input*, *information-processing*, *storage*, and *output*. Van Gelder and Port (1995) argue that these traditional information-processing terms do not adequately capture the nature of the relationships between body and world or of the processes that go on in these domains.

> The cognitive system does not interact with the body and the external world by means of periodic symbolic inputs and outputs; rather, inner and outer processes are *coupled,* so that both sets of processes are continually influencing each other. Cognitive processing is not cyclic and sequential, for all aspects of the cognitive system are undergoing change all the time. (p. 13)

> [A]ny fully adequate approach to the study of cognitive systems must be one that can handle **multiple, simultaneous interactive activity**. Yet doing this is the essence of dynamics. Dynamical systems *are* just the simultaneous, mutually influencing activity of multiple parts or aspects. (van Gelder & Port, 1995, p. 24, bold emphasis added)

So DST provides us with an appropriate concept for making sense of the frequent boundary crossings discussed in the previous chapter. Perception-action cycles can be understood in terms of the coupling of agents and environments rather than in terms of message passing (see Kugler & Turvey, 1987).

The Coupling of Reactive, Decentralised Architectures and the Environment Leads to Interactive Emergence

Dynamical systems explanations do not rely on any particular component of a system 'carrying information or content' about other aspects of the system or other systems that interact with the system. Rather, ordered behaviours are viewed as arising from the dynamic interactions between components of systems. In short, order is emergent in coupled dynamical systems (a product of interacting aspects) rather than enforced by a controlling device within the system. Hendriks-Jansen (1996) notes that:

> Dynamical systems theory has made it possible to conceive of complex behavior as arising interactively from the structure of the environment in conjunction with the creature's internal dynamics. We no longer need hierarchically organized planning systems to explain intricate temporal structure. A natural creature's behavior does not need to be preplanned. It does not have to exist as an abstract internal representation in the creature's head before it is "executed." The complex structure can emerge as and when it happens from the dynamic coupling between organism and its environment. ... [D]ynamical systems theory by itself cannot provide the natural kinds for an explanation, but it lays to rest the idea that complex behavior requires a computational explanation involving internal representations of the creature's activity that need to be prepared before they are "executed." (pp. 325-326)

The tools of dynamical systems theory provide a framework that can help us to understand how ordered activity can occur without a central orderer or controller. Order can simply

and reliably emerge from the interactions (often non-linear in nature) between multiple components. Beer (1995b) shows how the notion of coupling can be used to clarify the interactionist idea that behaviour is not pre-coded within the agent or the environment:

> Since properties of the coupled system cannot in general be attributed to either subsystem individually, an agent's behavior properly resides only in the dynamics of *A* or *E* alone. This suggests that we must learn to think of an agent as necessarily containing only a latent potential to engage in appropriate patterns of interaction. It is only when coupled with a suitable environment that this potential is actually expressed through the agent's behavior in that environment. (p. 132)

In chemistry, for instance, ordered patterns (systems with much reduced degrees of freedom), like the wave patterns in the Belousov-Zhabotinsky chemical reaction (see the next chapter), can emerge from a system that could potentially have many more degrees of freedom. In the Belousov-Zhabotinsky reaction there exist many millions of molecules of different chemicals that, under equilibrium conditions, would thoroughly intermix into a random, homogenous looking mixture. The chances that the chemicals in such a system would group like-molecules together, so that patterns would be apparent at the macroscopic level, are astronomically small (i.e., close to impossible). Yet, under non-equilibrium conditions (e.g., when heat is pumped into the system), such order *always* emerges (see Thelen & Smith, 1994).

If one were following a traditional cognitivist strategy for explaining the emergence of order in the Belousov-Zhabotinsky reaction, one might suppose that a qualitatively new set of instructions telling the molecules where to group (a new causal agent) had been introduced to the chemical system. Instead, what one finds is that the change from a disordered to an ordered regime is merely *triggered* by a steady change in a particular control parameter (e.g., temperature). "What is remarkable ...", argue Thelen and Smith (1994, p. 62) "... is that the parameter change ... [is] entirely nonspecific to $\boldsymbol{q}$ [the control parameter]. The temperature [has] no information whatsoever prescribing the nature of the chemical reaction ... . The pattern emerged strictly as a function of N, the nonlinear dynamics of the system." [97] (see also Kelso, 1995). Within an ESD framework similar analyses can be made of, for instance, the behaviour of situated robots. Instead of attempting to understand behaviour in terms of either the environment instructing the robot, or the robot containing instructions for action, a dynamical analysis can make

---

[97] In this example Thelen and Smith are using the variables in Haken's (1978) general dynamical equation to show how phase shifts due to a parameter change differ from 'instructional' commands by new causal agents:

$$\boldsymbol{q}=N(\boldsymbol{q}, \text{parameters, noise})$$

(where $\boldsymbol{q}$ is the behaviour of the system and N is the nonlinear function governing the elements of the system).

perfect mathematical sense of the idea that behaviour emerges from the interplay of environmental and agent-side resources.

Making Sense of Many Kinds of Change

Non-linear dynamical systems, as noted above, also have the natural facility for making sense of systems that change continuously (linearly, monotonically) within certain conditions and discretely (nonlinearly, non-monotonically) within others (see Elman et al. [1996] for a thorough analysis of different kinds of change). Often slight changes in a control parameter cause only slight changes in the phase portrait of a dynamical system – what you would expect from a simple linear system. But even simple nonlinear equations[98] that describe systems consisting of only a few variables and parameters related by simple rules of evolution can exhibit radical changes in the system's phase portrait. These phase transitions have all the appearance of events caused by new causal variables when in fact such events can occur as a natural consequence of the non-linear dynamics of the system. Within cognitive science this is a very important feature because computational stories of the structure and function of underlying mechanisms are largely driven by formal analyses of surface behaviour (including observed neurobiological and physiological patterns of activity). The computational theorist is thus vulnerable to introducing redundant complexity into their description of the underlying substrate. Sophisticated dynamical analyses are more likely to bring such possibilities to light.

Complex Causal Networks Rather than Simple Effective Causation

Perhaps most significantly dynamical analyses, especially analyses of non-linear, self-organising systems, show how it is possible to challenge the idea that the behaviour of agents is mechanically determined in a relatively simplistic manner by either external or internal forces (see, e.g., Juarrero, 1999). Even quite complex computational accounts of activity portray behavioural responses as the outcome of long, linear cause-and-effect chains. Yet relatively simple looking systems can exhibit self-organisation – that is, ordered transitions between global patterns in response to changes in environmental parameters. Such systems can be understood to have a kind of *natural and spontaneous*

---

[98] One of the most commonly cited nonlinear systems is known as the logistic mapping:

$$f(x) = rx(1-x)$$

By changing the control parameter r one can observe the behaviour of the system change from having a single limit cycle attractor (when r is less than 3) to having 2, 4, and 8 such attractors (at various values of r up to 3.54409). At r=3.569945 the system enters a chaotic regime returning briefly to having several limit cycle attractors occasionally at higher values of r (see Elman et al., 1996, pp. 214-219).

activity (they 'keep changing themselves') even in the absence of 'input'[99]. When such systems couple with other systems (including other self-organising systems) the coupling relationship is best understood as one in which the 'input' from the other system serves to *perturb* the dynamics of the system rather than to direct it how to act. Such systems do not always respond the same way to identical 'inputs' because the current internal state of the system affects how an input will be responded to. Thus complex systems are better thought of as compensating for perturbations and deformations than responding to commands.

> [E]ach of these two dynamical systems is continuously deforming the flow of the other (perhaps drastically if any coupling parameters cross bifurcation points in the receiving system's parameter space) and, therefore influencing its subsequent trajectory. Any agent that is going to reliably accomplish its goals in the face of such environmental perturbations must be organized in such a way that its dynamics can compensate for or even actively exploit the structure of such perturbations. (Beer, 1995b, p. 131)

Dynamical analyses of complex systems are perhaps more likely to provide us with scientifically satisfying explanations of the apparent autonomy of organisms than the traditional computational approach by maintaining basic scientific beliefs in determinism and causation.

Applying Dynamical Systems Theory to the Study of Cognition: Two Illustrative Examples

I will now illustrate the ways in which DST has been used to make sense of cognitive activity using two examples from the literature. Importantly, neither example involves an attempt to model real human or animal cognitive activity and this should serve as a cautionary warning to the reader (see Beer [2000] for a summary of some dynamical analyses of real-life examples of cognition). Both examples have been chosen because of the way in which they illuminate key DST principles at a fairly fundamental mechanistic level of analysis. Because of the structural complexities of real agents and real environments, current DST models of human and animal cognitive activity (e.g., Thelen & Smith, 1994) often do not exhibit a level of precision and fineness of detail necessary to convincingly demonstrate the important features of a dynamical approach. Thus, the following examples are simplified, artificial, and indeed, not even fully cognitive in nature.

---

[99] Importantly, complex, self-organising systems are, however, usually *open systems* – that is, systems which require a continuous supply of energy to maintain their integrity (which is a consequence of their far-from equilibrium nature). This sort of openness is not what I mean by 'input' which, here is taken to be of a more 'instructive' nature such as an introduction of a new chemical or the approach of a predator (see Thelen & Smith, 1994, pp. 54-56). In the next chapter I discuss Maturana and Varela's (1980, 1988) claim that certain kinds of complex system (an *autopoietic* system) are *structurally open* but *operationally closed.* This distinction maps on to my present distinction between 'receiving input' and being an open, self-organising system.

Hopefully, however, they should serve as adequate 'intuition pumps' for illustrating the promise of the DST approach to cognition.

Van Gelder's Analysis of the Watt Governor

Tim van Gelder (1995) provides one of the best known (and most widely criticised) examples of the ways in which a dynamical analysis of a quasi-cognitive problem differs from the traditional formal-task analysis approach of cognitivism (see chapter 2)[100]. He asks us to think of an 18th century technological problem – *the governing problem* – as being akin to a basic cognitive information-processing subtask such as recognition by matching or memory retrieval. Van Gelder asks us to imagine how a computationally-oriented researcher might try to solve the governing problem faced by engineers during the industrial revolution. During this period steam engines promised to provide a powerful and efficient workhorse for many industries including the cotton industry where spinning and weaving could potentially be automated and mechanised. However the speed of a steam engine's flywheel was known to fluctuate with changes in steam pressure and the workload being placed upon the engine. This lack of uniform engine speed meant that high quality spinning and weaving would be impossible under steam power. What was needed was some way of regulating the steam input into the piston (via the throttle valve) so that a consistent speed was maintained. Ultimately this problem was solved when James Watt invented the centrifugal governor.

Thus, the problem space set up by van Gelder's historical analysis was of the need for a quasi-intelligent physical device to adjust an engine throttle so that it would let enough steam into the piston to maintain a consistent flywheel speed. Think of the governor as a 'little person' behaving in the environment of the steam engine whose job it is to smoothly coordinate engine power output with workload.

As was noted in chapter 2, van Gelder suggests that the computationally-oriented engineer would resort to using a formal task analysis to understand the system. Recall that this involves breaking the overall problem task into smaller components in order to produce an algorithm that specifies the distinct 'information processing' steps carried out by the computational governor. This formal task analysis would look something like this:

1. Measure the speed of the flywheel.
2. Compare the actual speed against the desired speed.
3. If there is no discrepancy, return to step 1. Otherwise,
   a. measure the current steam pressure;
   b. calculate the desired alteration in steam pressure;
   c. calculate the necessary throttle valve adjustment.

---

[100] The governor example was used in chapter 2 to illustrate the use formal task description in cognitivism.

4. Make the throttle valve adjustment.
5. Return to step 1 (Van Gelder, 1995, p. 348).

The formal task analysis then demands that the computational governor that implements these steps should possess a collection of physical components for carrying out the various subtasks. These components would likely include "a tachometer (for measuring the speed of the wheel; a device for calculating speed discrepancy; a steam pressure meter; a device for calculating the throttle valve adjustment; a throttle valve adjuster; and some kind of central executive to handle sequencing operations." (p. 348).

What is of significance here is that the computational governor seems to face many of the problems that cognitivists claim humans and other animals face. These problems include perceiving and constructing representations of the environment, comparing these states of the world with desired (internal, knowledge) states, computing the differences between these represented states (presumably in some common form of computational calculus, e.g., a language of thought), working out what to do given these differences and then actually executing these 'motor programs'.

Watt's Solution to the Governing Problem

What is interesting, as van Gelder points out, is Watt's *actual* solution to the governing problem was nothing at all like that of our cognitivist engineer. Indeed, according to van Gelder, Watt's solution contained *nothing* that could effectively be understood as involving representations or computations and is better understood in terms of dynamical explanation. The implication that van Gelder and I wish to take from this is that perhaps a dynamical explanation is a better one for *all cognitive activity* and that the computational-representational model is at best a very rough approximation of some aspects of cognition and is at worst completely unhelpful.

Watt's solution to the governing problem consisted of building a device (a governor) that is attached to engine's main flywheel and to the throttle valve that regulates how much steam flows into the engine. The governor simply transforms the behaviour of the flywheel so that it adjusts the throttle valve in an appropriate manner. Van Gelder (1995) describes the structure of the governor in the following way:

> It consisted of a vertical spindle geared into the main flywheel so that it rotated at a speed directly dependent upon that of the flywheel itself … . Attached to the spindle by hinges were two arms, and on the end of each arm was a metal ball. As the spindle turned, centrifugal force drove the balls outward and hence upward. By a clever arrangement, this arm motion was linked directly to the throttle valve. The result was that as the speed of the main wheel increased, the arms raised, closing the valve and restricting the flow of steam; as the speed decreased, the arms fell, opening the valve and allowing more steam to flow. The engine adopted a constant speed, maintained with extraordinary swiftness and smoothness in the presence of large fluctuations in pressure and load. (p. 349)

So the governor's activity makes sense in terms of the physical mediation it performs between the flywheel and the steam inflow. It stretches the imagination to view such a process as a 'modelling' of the engine environment, a calculation of an appropriate response, and an execution of those calculated responses as classic cognitivist 'agent theory' would do. Now, as van Gelder notes, most people, engineers and non-engineers alike, gain a fairly intuitive mechanical understanding of how the governor accomplishes its task just by looking at how it operates as a piece of machinery. Moreover people never seem to try and understand the process in terms of parcels of information being sent from the flywheel to the governor where a representation of the engine speed, or current workload, is created in the structure of the governor (e.g., in the arm angles) which is transformed in order to produce a motor program for instructing the valve to open and close. Dynamical systems theory provides a precise mathematical way of formalising the intuitive mechanical understanding. With the arms attached to the flywheel, but not to the valve, the governor's dynamics can be described by the following differential equation:

$$\frac{d^2\theta}{dt^2} = (\eta\omega)^2 \cos\theta \sin\theta - \frac{g}{l}\sin\theta - r\frac{d\theta}{dt}$$

The parameters represented in this equation are current engine speed (ω), the gearing constant (η), the gravity constant (g), length of the governor's arms (l), and a friction constant (r). The only varying quantities are time (t) and the state variable of arm angle (θ). The equation thus tells us how the 'change in arm angle' changes (the instantaneous acceleration) depending on the current arm angle, the current arm angle change, and the engine speed. To find out, for instance, the angle of the arm at time t one would have to find a solution (i.e., another equation) to this general equation.

The dynamics becomes more complicated when the governor is hooked up so that it regulates the engine speed, as we require another equation representing the dynamics of the steam engine. Van Gelder uses the following differential equation to describe the dynamics of a steam engine that might be hooked up to a governor.

$$\frac{d^n\omega}{dt^n} = F(\omega,...,\tau,...,)$$

In this equation τ is the parameter for the current state of the throttle valve (which depends directly upon the governor's arm angle θ) and ω represents engine speed. This equation

tells us how change[101] in engine speed changes as a function of engine speed, throttle valve state, and whatever other relevant variables one might need to take account of in an equation describing the dynamics of a steam engine. When coupled with the governor equation (i.e., when the values for ω and τ [in terms of θ] are shared), the entire dynamics of the overall governor-steam engine system is described.

In an important sense the mathematical equations are not of primary importance here. They merely serve to show that a formal non-representational calculus does exist for making sense of the coupled activity. The key difference between the computational governor and the Watt governor is in the contrast between thinking of the behaviour of the system in mechanical terms rather than computational and representational ones. However, the dynamical analysis does accentuate the following aspects of the governor-steam engine system:

1. The governor and the steam engine each have their own intrinsic, ongoing and quantitative, dynamics. This does not necessarily mean that the two subsystems can operate independently of each other. In this case the steam engine can obviously function without a governor, but the governor will not spin without energetic input from the engine. Nonetheless the governor's structure specifies a space of possible behaviours (i.e., a phase portrait) independent of any particular 'input'.
2. When coupled, the governor and the engine exhibit simultaneous, quantitative, modulatory activity. One does not simply impel the other to work. They are not static entities that require the other to send them information in order to act. They each have their own intrinsic dynamics (represented by a phase portrait) which is perturbed or parameterised by the other.

> [A]ny change in engine speed, no matter how small, changes not the state of the governor directly, but rather the way the state of the governor changes, and any change in arm angle changes the way the state of the engine changes. (van Gelder, 1995, p. 357)

Van Gelder's examples provide two lessons. The first lesson is that the traditional strategy of formal task analysis sets us off in a direction where we build (in the case of simulation – a synthesis problem) or model (in the case of cognitive analysis – an analysis problem) *componential systems* (or *homuncular systems*) where the overall activity of a system is accomplished by the aggregation of the activities of its subcomponents (see Clark, 1996, 1997). Van Gelder implies that a DST approach does not constrain our thinking in this way and indeed may, by the mere fact that the framework is so *un*constraining, alert us to the

---

[101] This equation is a rather vague and general one and meant only as a sketch. Van Gelder's (1995) use of the index *n* indicates that he is only concerned with discussing *some kind* of derivative of engine speed. Thus my use of the term *change* here is used in a similarly vague sense.

fact the there may exist non-componential solutions to our analysis and synthesis problems. One has to be cautious here, however, and note that something more than 'lack of constraint' is necessary to get at non-componential (i.e., emergent and dynamical) systems including possibly the use of behaviour languages (Brooks, 1997), artificial evolution (Beer, 1995a, 1995b; Harvey et al., 1993), and ethological investigation (Hendriks-Jansen, 1996).

The second of van Gelder's lessons is that cognitive systems are better understood as dynamical systems than as traditional representational-computational ones. This claim is just a statement of the dynamical hypothesis. He argues that "the deepest reason for supposing that the centrifugal governor is not representational is that, when we fully understand the relationship between engine speed and arm angle, we see that the notion of representation is just the wrong sort of conceptual tool to apply." (van Gelder, 1995, p. 353). For van Gelder the operation of the governor is just too subtle and complex for the traditional notions of representation and computation to handle. Digital computation is essentially sequential and cyclic in nature and implies the existence of discrete, identifiable steps in the process. Yet the coupling of the governor and the engine is a smooth, continuous, quantitative, and simultaneous process. DST incorporates tools that can capture these subtleties and complexities.

While van Gelder's arguments have been influential in the first case, a number of authors have questioned his second claim. In particular these authors have been at pains to argue that some interesting sense of representation can still be gleaned in the governor example. Van Gelder argues that the angle of the governor's arms cannot be understood as a representation of the current speed of the engine (pp. 352-353)[102]. Bechtel (1998), Chemero (1998a, 1999), and Markman and Dietrich (1998) argue that it can.

Problems with Computational Readings of Watt's Governor

As often happens with attempted non-computational explanations of cognitive activity, unorthodox theorists need to be wary of orthodox theorists claiming that the unorthodox alternative is just a particular implementation of a computational-representational system. Within modern cognitive science, classical symbolicists contend that connectionist

---

[102] Interestingly van Gelder has been widely misinterpreted as arguing that *all* dynamical systems theoretic approaches to cognition must be anti-representationalist in nature. However, he clearly states that DST is probably compatible with thinking of representations as, perhaps, attractors, parameter settings, trajectories, or system states in some systems (see van Gelder, 1995, p. 376; van Gelder, 1998, p. 622; van Gelder & Port, 1995, pp. 11-12). He *does* explicitly argue that the centrifugal governor example is non-representational in nature and this has worried a number of representationalists. Perhaps this is because they see similarities between the governor example and cognitive phenomena that they are used to thinking of as representational.

networks are just a particular kind of symbolic system. The same fate seems to await the excited dynamicist (Markman & Dietrich, 1998; Bechtel, 1998; Chemero, 1998, 1999). Van Gelder is alert to this possibility and points out that there are several significant aspects of the Watt governor that simply do not fit the cognitivist's story.

At first glance it may appear that the angle of the arm on the governor represents (in the sense that the arm's position reflects or covaries with) the speed of the engine. Van Gelder (1995, pp. 352-353) argues that this is not the case because: 1) the mere fact of covariance does not make something a representation (see my critical discussion of covariance theories of content determination in chapter 3); 2) that in any event the arm angle hardly ever correlates with engine speed because the governor reacts much more slowly to extra loading on the machine than the engine does – that is, at times when the whole system has reached a stable equilibrium point, the arm angle follows one kind of 'correlational scheme' while at other times (when the workload on the engine changes or the engine's pressure drops) it follows another; and 3) a representational analysis of the role of the governor's arms is simply wrongheaded. The point of the *arms,* as part of the entire governor assembly, is to enable a balanced output of power not to signal to the valve mechanisms what the current state of the engine is. Rather, both subsystems – the governor and the engine – modulate each other's activity in a complex codeterministic manner. It is in fact important that the arms do not correlate with engine speed directly or even with a particular temporal delay because to do so would jeopardise the proper functioning of the engine-governor system[103].

Bechtel (1998, pp. 301-306) argues against van Gelder's dismissal of the governor's arm angles as representations of engine speed. Bechtel takes something to be a representation in the sense that I outlined it in chapter 2. That is, a representation is an object or event that carries information (a bearer) about another object or event (a referent) that is used by a system (a user) in order to coordinate its behaviour with the referent. So Bechtel, following Haugeland (1991), takes something to be a representation if it *stands in for* the object it represents within the operations of a cognitive system.

Against van Gelder's first point, he argues that the governor possesses all of the criteria necessary for counting the arm angle as a representation of engine speed. That is the arm

---

[103] Smithers (1994 cited in Clark, 1997, pp. 95-96) notes that the second generation of governors were designed to react more quickly to engine speed changes due to factors such as extra workload by using better machining techniques that cut down on friction between parts, but that their increased sensitivity led them to malfunction by continually 'hunting' (oscillating between slowing down and speeding up). In order to overcome this problem the new governors had to incorporate components to avoid hunting. Thus we see that the perfect correlation beloved of causal theorists of representation (e.g., Dretske, 1995) actually proves to be a liability in this case!

angle is designed so that the 'information' in the arm angle can be used to modulate engine speed – it has a user, a bearer, and a referent. This, so Bechtel argues, effectively answers van Gelder's 'mere correlation argument'.

Bechtel deals with van Gelder's second objection by suggesting that the time-lag between the state of the engine and the 'relevant' arm angle does not necessarily preclude the arm angle being a representation. He supports this claim in two ways. First he marshals Millikan's biological function argument (see chapter 2) that x may represent y even if the two never correlate. Second he argues that an effect (in this case the state of the representation bearer) is always going to lag behind its cause (in this case the referent) making the two always temporally 'out of sync'.

Finally Bechtel takes on van Gelder's claim that the representational relation is the wrong sort of way of understanding the mutual modulation of governor and engine dynamics. He argues that just because the governor and the engine are related in a subtle, complex, and mutually determining way, "it is not clear why it is too subtle and complex to satisfy the stand-in aspect of representation. Something can stand in for something else by being coupled in a dynamical manner, and by being so coupled figure in determining a response that alters the very thing being represented." (p. 304).

I believe that Bechtel's points largely miss the mark even though they raise some interesting points about a possible way of 'dynamicising' representational approaches. First of all, it is not clear that Bechtel fully appreciates the nature of the non-correlation between arm angle and engine speed. This lack of correlation involves not just a simple temporal delay but in fact a distinct lack of *overall* covariance. So it is not clear how representational content could be recovered from this more complex relationship between arm angle and engine speed. It is also not clear how Millikan's (1984, 1993) teleological approach would make much sense of what is going on. Chemero (1998, pp. 102-104) attempts to do just this, but the analysis is much less illuminating than that provided by the dynamical perspective (see the following section on a dynamical analysis of a Sussex robot for Chemero's re-analysis of the second example to get a feel for what a teleological account gives us). In an important sense Bechtel's efforts are primarily aimed at saving, not a representational account of agent-environment systems, but a *mechanistic* account (I discuss this issue in more detail in the final section of this chapter). His main concern is with DST's ability to give us an account of the target system that will illuminate the important causal relations that hold between its components. It is not clear to me why a representational analysis falls out as a natural consequence of holding the belief that a componential analysis is required to supplement the insights of a dynamical analysis. But this seems to be what Bechtel (and Clark, 1996, 1997) argue. There are a number of possible reasons why this representational assumption is regularly made. Here are two:

1. The first is that the notion of internal representation continues to evolve. With the advent of the notions of *action-oriented representations* and their kin (see the previous chapter and Clark, 1997; Chemero, 1998a), it has become permissible to think of structures that simultaneously 1) deform systematically in the face of specific environmental structure and 2) control/modulate motor activity, as representations, although it remains to be seen how content is handled in these cases. So the fact that the governor's arms may signify both the speed of the engine and the change in the valve opening makes perfect action-oriented representational sense. Although the notion of representation is becoming increasingly refined by exposure to ESD interactionist research, it is also becoming increasingly removed from its roots as an off-line model or simulation of absent objects, properties, or events (Keijzer, 1997). Just because examples such as van Gelder's can be reworked to accommodate some kind of stripped down, modified notion of representation, it does not follow that representational accounts provide the best or most useful accounts of cognitive activity for analysing, understanding, and modelling the behaviour of cognitive systems.

2. The second reason for the ongoing debates between representationalists and non-representationalists is that it is probably the case that the governor example is just too simple and too un-cognitive to ram the DST point home. Many mechanical systems succumb to componential/representational analyses quite easily because they have been constructed according to componential principles (Clark, 1996; Hendriks-Jansen, 1996). Although the real centrifugal governor is not built like van Gelder's fictitious computational governor, it has still been designed as a component for modulating the activity of another machine. So perhaps it should not be surprising that some sort of representational gloss can be given.

The second example provides a more complex, more 'natural' (because it features a less explicitly 'designed' agent), and more distinctively cognitive example of the ways in which a dynamical systems explanation can be deployed in the study of agent-environment systems. Once more I will use an example from the situated robotics literature, this time from the work of the evolutionary robotics team of Inman Harvey, Phil Husbands, and Dave Cliff (HHC hereafter) at the University of Sussex (Cliff et al., 1993; Harvey et al., 1993, 1994; Husbands et al., 1995).

A Dynamical Analysis of a Sussex Robot

As noted in previous chapters, HHC have designed and built a number of artificially evolved mobotic systems. Genetic algorithms are used to select connectionist architectures that produce simple adaptive behaviours such as target following and centering the robot's position in a room. HHC have provided dynamical, rather than representational analyses of these robots. Wheeler (1996) argues that the resulting connectionist architectures push our

attempts at giving a representational gloss to these behaviour-producing systems to the very limit.

Husbands et al. (1995) argue that the operation of these robots in their environments is best understood from a dynamical systems perspective. They adopt a similar position to Beer (1995a, 1995b) in arguing that computers (computational systems) make up only a small proportion of all possible dynamical systems[104] and that "the all-too-common assumption that cognition is some form of computation is ... a stultifying restriction on possible models for a robot control system, as well as the rest of AI." (p. 85).

HHC understand the activity of their evolved robots in terms of the coupling of an agent system and an environment system in the same sort of way that van Gelder does with the governor-steam engine example. For the agent-side part of the dynamical equation Husbands et al. (1995, pp. 91-93) describe the robot's internal control system using equations that represent the connectivity and connection weights within the robot's neural network[105] as well as mathematical descriptions that describe the characteristics of the robot's visual sensors (two low bandwidth photoreceptors that register brightness).

The robots studied by HHC are 'evolved' in a simple environment consisting of a cylindrical 'room' with walls of varying heights. This environment side of the dynamical equation is described by Husbands et al. (1995, pp. 93-95) in terms of the kind of 'input' the robot receives from at various regions in this environment. The input is described in a

---

[104] HHC use a quite general characterisation of a dynamical systems as "... any system that can be characterised by a finite number of state variables, and a dynamical law that specifies how these state variables change with time." (Husbands et al., 1995, p. 85). This is a much looser and more inclusive definition than that used by van Gelder (1998). He argues (pp. 617-619) that dynamical systems are *quantitative systems* – that is, systems whose behaviour exhibits quantitative distances between states. However, in digital computational systems, such as Turing machines, states are merely ordered. This distinction makes the dynamical hypothesis a non-trivial claim (see Wheeler's, 1998 commentary on van Gelder for more detail).

[105] Husbands et al. (1995, pp. 87-89) utilise recurrent dynamic real-time networks. These networks are more complex than traditional three-layer, feedforward networks. Each network consists of a set of preset 'input' nodes and a set of preset motor controlling nodes as well as a set of internal nodes. The artificial evolutionary process can connect nodes in any manner (forward, backward, direct feedback to self, or unconnected). Moreover, any number of internal nodes can evolve – that is, some robots may produce 5 internal nodes, some more, some less – although there are only fixed numbers of input and output nodes. Interestingly, several of the evolved designs 'cannibalised' unused input nodes and used them as internal nodes and some output nodes were not used (i.e., remain unconnected). So, even this fixed input and output architecture was effectively plastic. Finally, the connections between nodes incorporated time-delays (unlike traditional networks where all nodes are updated simultaneously) and an element of random noise. This led to a number of units forming noisy feedback loops – essentially ongoing self-organizing activity. This activity was perturbed but not caused by input.

robot-relative manner in terms of a space consisting of two parameters: r being the point on a radius that passes through the centre of robot and that originates from the centre of the cylindrical environment, and ϕ being the clockwise angle between the radius and the 'front' of the robot. Motion (moving in straight lines, turning left and right etc.) is described in terms of transformations in this relativistic coordinate system. So r and ϕ evolve according to dynamics of the robot's movements.

The overall activity of the robot is described in terms of the coupling of these two systems. Simply put, the agent subsystem consists of a set of variables (activation levels of units, connection weights, etc.) that evolve according to a certain dynamic (the linear threshold function that 'sums' inputs into units). The parameters consist of two environmental 'input' values (*Er* and *El*) corresponding to the input registered by the robot's photoreceptors. These parameters are variables within the environment (or stimulus array) subsystem which 'evolves' according to transformations produced through the movement of the robot. HHC summarise this overall activity in terms of the vector field (phase portrait) of the entire system U (see Husbands et al., 1995, pp. 97-102 for this global analysis).

Wheeler (1996, pp. 218-225) argues that HHC's dynamical analyses provide us with a better explanatory and predictive tool of the robot's activity than a representational-computational approach can because a dynamical account can better address the ways in which activity emerges from the interplay of the robot's network and sensors and the layout of the environment (i.e., evolved animate vision in action). He focuses on the facts that 1) the robot manages to carry out relatively sophisticated navigational tasks using very simple sensors, and 2) classical information-processing models do not encourage us to think that such things are possible.

> Minimal monocular vision is no basis for building models of the world, which then could be used to make decisions about how to proceed. The robot succeeds in its task by exploiting its own movements to create variations in light inputs. Thus the ways in which the sensory-motor mechanisms are coupled to the world enable the robot to complete the task, given its ongoing activity as a whole agent. (p. 224)

He goes on to claim that the Cartesian view, as he calls it, guides theorists to think of solutions to cognitive problems purely in terms of what is inside the agent, whereas a dynamical coupling approach facilitates understanding how activity can interactively emerge from the interplay of a moving agent and its surroundings.

> The visuomotor phase portraits, which were constructed to understand the behaviour of the room-centring robot in different environments, could be thought of as an attempt to capture how the robot moves through its visuomotor world. The dynamical structure of that world (the phase portrait) changes when the robot's 'physical' environment (to which it is coupled) changes, although the structure of the control system itself does not change. (p. 225)

Wheeler's conclusion is akin to the claims of van Gelder's (1998) knowledge hypothesis that cognitive activity is best *understood* using the tools of DST rather than necessarily reflecting the actual mechanisms underlying the robot's behaviour (the nature hypothesis). However, Chemero (1999) disagrees with Wheeler's analysis, arguing that it *is* possible to formulate an illuminating representational analysis of HHC's robots. He analyses the roles of the various nodes within the network of one of HHC's evolved target-following robots (Harvey et al., 1994) and provides a representational analysis based upon his favoured teleological theory of representation. The robot's final network structure consisted of 12 nodes[106] connected by a variety of feedforward and feedback (recurrent) excitatory and inhibitory connections. Two nodes (0, 1) served as input nodes from two visual detectors (*v1* and *v2*). Three other nodes (13, 14, 15) served as output nodes that controlled the motor and steering capacities of the robot (13 increased voltage to the left wheel motor, 14 decreased voltage to the left motor, and 15 increased voltage to the right motor). The other seven nodes served as intermediaries between input and output with the exception of node 5 which excited other input and output nodes but received no input itself.

Chemero uses Harvey et al's (1994) analysis to make his representational claim. The latter found that the robot exhibited three kinds of behaviour each of which was controlled by a subnetwork of units (akin to the behaviour layers in Brooks' mobots):

1. When the signals to both *v1* and *v2* were weak (i.e., when the light coloured target was not 'in view') little or no activation was propagated to the input units. This resulted in the noisy node (unit 5) have primary influence over the network's activity which in turn caused the robot to turn on the spot in an irregular, noisy fashion. This provided the robot with the opportunity to 'wander about' until a strong input was detected in either *v1* or *v2* thereby setting off a different kind of behaviour.
2. When *v1* received a weak signal and *v2* a very strong one, unit 1 self-inhibited and triggered a pattern of activation that led to the same kind of noisy turning as in the previous condition. However, when *v2*'s signal was "medium high" (Harvey et al., 1994, p. 399) a pattern was set up that caused unit 14 to excite itself so that the left motor slowed. This caused the robot to rotate in a medium radius circle until a strong signal was detected in *v1* and the following conditional behaviour occurred[107].

---

[106] The artificial evolutionary process led to a number of other nodes being effectively unconnected to the network. The 12 nodes mentioned here correspond only to those that played a role in the robot's behaviour production.

[107] Chemero (1999, pp. 11-12) seems not to distinguish between the different behaviours caused when *v2* receives strong input and when it receives medium high input. It appears that he bases his analysis on the medium high input condition. Although this does not weaken his basic argument, it provides an impression

3. When *v1* received strong input, regardless of the strength of the input to *v2*, unit 0 triggered a pattern of activation which led to both motors being activated. This caused the robot to move straight ahead.

Together these behaviour layers resulted in the strategy of rotating until the target was detected in *v1* and then of moving toward the target in a straight line. If the target was ever lost or obscured the robot would again rotate until it was acquired by *v1*.

Chemero argues that this set-up is enough to warrant branding the system as representational according to his teleological theory. The input nodes correspond to *representation producers*, the output nodes are *representation consumers*, and the patterns of activation across the intermediate nodes are *representations*. Under a teleological analysis the contents of the representations within the robot correspond to *the way the world would need to be for the behaviour caused by the representation consumer to be adaptive*. So in condition 1 the pattern of activation of the intermediate units have the content of, roughly speaking, **target out of sight**. In condition 2 they have the content **target in sight (but not straight ahead?)**, and in condition 3 the content corresponds to **target ahead**.

Yet despite successfully producing a representational gloss for the operation of the Sussex robot, Chemero argues that representationalists have little to celebrate. He gives several reasons why we should prefer the *dynamical* explanation over a representational one (Chemero, 1999, pp. 14-15).

First of all Chemero notes that his construction of the representational story depended upon the prior existence of the dynamical account (p. 14). In a sense he used the dynamical account to run through the architectural (network) implications of placing the robot in different positions in the environment. This gave him the information necessary to assign 'roles' or 'content' to different network node activation patterns. This makes the representational account an *a posteriori* one. Moreover, since Chemero focuses on only one of the robots evolved by HHC, we do not get a chance to see the advantages of the dynamical account in dealing with a collection of different but related control systems – including 1) systems with different architectures evolved to deal with the same task and 2) systems evolved to deal with different tasks in similar environments. Whereas a representational analysis will require the theorist to treat each variation pretty much as a completely new problem to be analysed, a dynamical account may only require the relatively simple re-valuing of a couple of parameters. For instance, a variant network architecture can be dealt with by the dynamical analysis by 'simply' plugging in the new

---

that the robot's workings are simpler than they actually are and thus it is possible that this simplification works in favour of his representational argument.

network dynamics into the relevant parts of the coupled equations, provided the basic kind of network is not very different from the original. Provided one has the software available to plot the new phase portrait or vector field, and this is pretty much a pre-requisite for any dynamical analysis, the new description of the robot's behaviour is instantly available. In non-linear dynamical equations, such as those that exist in most interesting neural nets, small changes to an architecture or a task often result in the new robot exhibiting quite different behaviours from its predecessors. But these large changes can be instantly had by twiddling the relevant equations in a dynamical analysis.

By contrast, in a representational account these large changes in behavioural strategy would form the basis for the analysis. Representational accounts, of course, usually rely upon the researcher observing the behaviour of an agent, working out what stimuli it is reacting to, and then postulating how the agent is 'processing' and using that information. Because content assignment, the essence of any representational approach, relies on matching internal activity with behavioural and environmental kinds, representational analyses must exhibit a linear relationship between the behavioural complexity of the agent and theoretical complexity of the analysis of its behaviour. Of course, representational analyses may possibly be automated in simple cases, but this does not change the fact that a representational analysis is a much more labour intensive activity.

Secondly, and perhaps more importantly, Chemero argues that the representational account adds nothing to the existing dynamical analysis (van Gelder [1995] makes a similar point about the possible accounts of the operation of Watt's governor). With the mapping of the system's phase portrait we can successfully predict the consequences for the entire system given any set of initial conditions. I would even go so far as to say that the dynamical account is *more* useful than the representational one. Representational explanations typically do not give us the predictive power of a complete dynamical account. For instance, Chemero's vague reference to *patterns of activation* belies the complexity of the robots' networks and the numerous parameters (e.g., connectivity, noise, weights) that must be calibrated in order to produce the appropriate behaviours. Certainly one could not build or model in detail the inner workings of these robots using the representational analysis provided by Chemero.

Of course, HHC's robot system is relatively simple and we are unlikely to be able to produce anything like as precise a story for the behaviour of real agents in complex environments. We must guard against overextending our explanatory enthusiasm. However, it is important to note that HHC's robots are not all *that* simple and that it is hard to see that a traditional representational analysis of this system would have anything like the predictive power of the dynamical account. Yet demonstrations such as these do not convince the determined representationalist who argues that a representational explanation gives us something more than a dynamical one. It is to these issues that I turn next.

Does the Dynamical Approach Provide a Genuine Explanatory Framework?

There are two basic kinds of objection to the dynamical hypothesis: 1) that the application of dynamical systems tools to cognitive activity is unworkable, and 2) that dynamical analyses do not provide us with an explanation that is fertile enough for understanding cognitive mechanisms at a level adequate for a sophisticated cognitive science. I will examine each of these objections in turn.

Dynamical Systems Theory is Probably Unworkable in the Context of Cognition

A number of writers have observed that the tools of dynamical systems theory work best when the researcher is only dealing with a few variables (e.g., Clark, 1997; Eliasmith, 1996). It has been claimed that, beyond perhaps five to ten variables, the complexity of a dynamical system begins to threaten our "intuitive geometric understanding" (Clark, 1997, p. 101) and problems become intractable. Eliasmith (1996) argues that solving dynamical equations with the number of variables that, say a middle-sized connectionist network possesses, will prove to be impossible[108]. Such objections are probably a little premature. Husbands et al. (1995) happily construct dynamical descriptions of their quite complex neural network-controlled robots. Beer (1995a) points out that a savvy theorist can analyse high-variable systems by carefully picking and choosing which variables to analyse. Moreover these sort of objections do not challenge the contention that cognitive agents may well be dynamical systems (the nature hypothesis) in the sense of systems made up of quantitative variables (Van Gelder, 1998).

Dynamical Systems Theory needs to be Complemented with a Mechanistic Emphasis

A number of critics have remarked that dynamical systems explanations seem to leave out important aspects of the full explanatory project of cognitive science. In this section I want to take a quick look at some of these objections and pull out what I think are the most important points.

A common objection to, or at least misgiving about, the use of a dynamical systems framework in the explanation of cognitive activity is that it merely provides an abstract

---

[108] It is not clear in Eliasmith (1996) whether he fully appreciates what it means to *solve* a differential equation. This is an entirely separate task from the relatively simple task of specifying an evolution equation for a set of variables. Even some very low variable problems (such as the famous three body problem) are computationally intractable (unsolvable analytically) while the behaviour of many systems possessing a large number of variables can easily be modelled, that is, have their evolution equations described. The number of variables in a system is not the only measure of complexity. There exist a number of modern, often computer-intensive, methods for accurately approximating solutions for many analytically intractable problems including using difference equations (see Stewart, 1997).

mathematical description of agents' behaviours and that a complete theoretical explanation will also require a description of the underlying causal mechanisms, such as neural circuits or mental modules, which give rise to the surface behaviour described by dynamical equations. As I have noted at the beginning of the chapter, this claim is often, but not always, taken to be identical to the argument that a full explanatory theory of cognition must include reference to representations and computation. Van Gelder (1998, p. 625) is dismissive of such objections. He argues that other scientific endeavours, such as celestial mechanics, accept DST explanations as sufficient for all scientific purposes and that cognitive scientists should follow suit[109]. He suggests that cognitive scientists have traditionally adopted a relatively peculiar explanatory approach in utilising functional analyses of the components of 'cognitive systems', and that they have "become so accustomed to such explanations that anything else seems inadequate." (p. 625).

In contrast to van Gelder, both Bechtel (1998) and Clark (1997) argue that dynamical systems theory has its place is a modern cognitive science, but that its use does not preclude the need for traditional mechanistic and computational-representational explanations. Bechtel suggests that dynamical systems explanations are a form of covering law explanation popular in the logical positivist philosophy of science pioneered by the likes of Carnap, Nagel, and Hempel (Bechtel, 1998, pp. 306-307). According to Bechtel, covering law explanations take a phenomenon to be explained "when a statement describing it was derived from statements specifying one or more laws and relevant initial conditions." (p. 306). He argues that such explanations are better suited to the physical sciences (and some would argue that they are not even appropriate there) and are rarely found within the life or cognitive sciences. By contrast cognitive science, amongst other sciences, makes use of what he calls *mechanistic explanation*, or what Clark (1996, 1997) calls *componential* or *homuncular explanation*. Such explanations analyse the system in question in terms of the roles played by its components. The behaviour of a system is best

---

[109] Clark (1997, p. 117) points out that DST does supply an *explanation* rather than a mere description insofar as it provides a framework for understanding counterfactuals – the idea that a dynamical systems equation will systematically tell us how the behaviour of the system would be different if some variable of the system was changed. An inductive generalisation by contrast will not tell us anything about counterfactuals. In a similar manner Chemero (1999) makes an interesting case for *not* thinking of DST explanations as being purely superficial descriptions of surface activity and thus being a weak guide to discovery – a claim made against the phenomenalist physics of Mach by atomists such as Boltzman. Those physicists that posited hidden causal mechanisms (e.g., atoms) were able to make novel (and ultimately fruitful) empirical predictions about various physical phenomena. Chemero points out that it is possible that DST explorations of cognition may result in the discovery of dynamical equations that generalise to many kinds of behaviour. He gives as an example Kaipainen and Port's extension of the Haken-Kelso-Bunz model of rhythmic behaviour to a possible *general theory of meter* – a model that can be used to model any kind of rhythmic human behaviour (e.g., arm swinging, walking, and even speech actions).

understood by 1) thinking of the overall behaviour of the system as the outcome of many sub-behaviours (or the 'execution of component tasks' as Bechtel puts it), and 2) by assuming that each of these subtasks is accomplished by a separate (physical or functional) component of the system. Bechtel (1998, see also Bechtel & Richardson, 1993) refers to these two assumptions as the *decomposition heuristic* and the *localisation heuristic* respectively. Such a style of explanation focuses upon the internal structure of the 'behaving system'. Machines, such as washing machines, computers, and so on, succumb to this kind of functional analysis quite easily for the simple reason that many modern machines are designed using the 'reverse application' of the decomposition and localisation heuristics (see Hendriks-Jansen, 1996).

Thus, Clark (1997) does not believe that a dynamical systems approach is enough for cognitive science. He follows Dretske (1994, cited in Clark, 1997) in arguing that a complete explanation will give us the capacity to *build* the system being analysed; at least in principle. In order to do this one needs a full causal-mechanistic analysis of the system. And, since dynamical analyses often use relatively few state variables that correspond to quite abstract states of affairs, dynamical explanations do not provide the theorist with the knowledge needed to understand the causal-mechanistic links that hold between the components of the system. Clark (1997) claims that:

> [T]hese "pure" models do not speak directly to the interests of the engineer. The engineer wants to know how to build systems that would exhibit mind-like properties, and, in particular, how the overall dynamics so nicely displayed by the pure accounts actually arise as a result of the microdynamics of various components and subsystems. ... What is really being suggested is not that in fact that we should be able to build systems that exhibit the desired features ... but that we should understand something of how the larger-scale properties are rooted in the interactions of the parts. (pp. 120-121)

Keijzer (1997, pp. 195-198; see also Keijzer et al., 1998; Keijzer and Bem, 1996) makes a similar claim. Unlike Clark, who argues that a mechanistic explanation adds some explanatory power by telling us how the parts of the target system interact, Keijzer's interest in constructing a mechanistic level of analysis is in finding out how to apply the dynamical analyses in the first place. Without having what Keijzer calls a *theory of the implementing substrate,* one cannot even specify the relevant parameters and variables which would feature in our dynamical equations (Thelen & Smith [1994] make a similar claim). Keijzer argues that a developed interactionist theory of behaviour can and should be constructed relatively independently of any dynamical modelling. What we cognitive scientists want to know first is what sort of *organisation* underpins a system that is capable of adaptive behaviour. We need to be able to say, at a mechanical level of description, what separates the cognitive from the non-cognitive. Dynamical systems theory can not do this by itself. I believe that Keijzer is entirely correct in his analysis but that it may potentially threaten the radical insights that a dynamical approach can provide because it is sometimes assumed that the insufficiency of DST as a theory of cognition means that we *must*

embrace a representational-computational framework to fill in the explanatory holes. For instance, both Clark (1996, p. 264, 1997) and Bechtel (1998) argue that a system that is successfully understood in a mechanistic and componential manner lends itself to a representational interpretation because "what must be done in such accounts is to explain how information is carried through the system and made available to other parts of the system that use it." (Bechtel, 1998, p. 313). In other words, a representational analysis is thought to be pretty much identical to a componential mechanical analysis. Any system where we can think of components signalling other components is (somehow) automatically representational in nature.

Hooker (1997), however, argues that representational analyses typical in cognitive science do not usually tell us anything particularly interesting about the mechanical structure of a system. Ironically, the representational approach is actually *more* at risk of providing abstract redescriptions of surface behaviour than the dynamical account according to Hooker. Indeed, in an important sense, it is one of the dynamical approach's strengths that it requires a description of components in a different vocabulary to the one that describes the high level activity patterns of the system. Hooker (1997) claims that

> [S]tructuralist theories specify structures in the same cognitive terms as the cognitive behaviours they wish to explain – if that is an explanatory weakness, it has the advantage of not requiring explanations in sub-conceptual terms. By contrast, precisely because the dynamical systems approach wishes to explain the formation of global order by underlying interaction processes among more local components, its description of those components cannot be at the same conceptual level as the global order to be explained. (p. 107)

Thus, Hooker effectively turns the argument of Clark and Bechtel on its head. The cognitive architectures that cognitivists derive from formal task analyses are often abstract algorithmic redescriptions of the surface behaviour being examined (see also Hendriks-Jansen, 1996)[110]. Representational (structuralist) analyses are more like 'mere redescriptions' than DST accounts because the latter demands a complementary mechanistic analysis, whereas the former only *seems* to provide one.

I think that many of these criticisms of dynamical analyses of cognition raise valid points. A full explanatory account of a cognitive activity will require some causal-mechanical understanding of the ways in which a cognitive system (whatever that might be) is put together. And one cannot get this merely by using the tools of dynamical systems theory. It is important to point out, however, that the dynamical analyses of the likes of Thelen and Smith (1994; Thelen, 1995) and Saltzman (1995) utilise empirical analyses of behaviour,

---

[110] It is possible that the strategy of using formal task descriptions could be wedded to a set of precautionary principles that offset the straightforward equation of organisation of cognitive mechanisms with behavioural redescriptions. I suspect that development of the subsumption architecture concept (Brooks, 1991a) is partly the result of this kind of cautionary use of 'functional analysis'.

as well as physiological and neurobiological data, in hunting out the variables and parameters that feature in their dynamical accounts. Interestingly, many of these theorists do *not* use the language of computation and representation to make sense of these causal-mechanistic phenomena, contrary to the claims of Clark and Bechtel that such an analysis is a concomitant of an information-processing vocabulary. It is this aspect of many of the cognitivist arguments against DST that I think is unnecessary. It is important to be clear what I am objecting to here in these arguments. It is not the idea that components can be understood as signalling each other per se. I have no particular problem with the idea that one may want to try and understand the operation of, say, a washing machine, in terms of the way different components signal each other to operate. After all this sort of thing is exactly what is going on inside Brooks' (1991a) mobots all of the time. Rather, what I believe is problematic and ultimately unnecessary is the more specific claim in the study of cognition that these component-to-component signallings contain content about the things in the world outside of the system[111]. And this after all is the core notion of cognitivism: that at a subpersonal level mind/brains model the world in some important sense. The task for the ESD interactionist is to sketch out a broad causal-mechanistic perspective that could underpin a non-representational interactionist approach to cognition. That is task I tackle in the next chapter. For now, however, it is important to summarise the moderate interactionist attitude to DST that has been developed in this chapter.

What does Dynamical Systems Theory do for us?

Dynamical systems theory gives us the tools for describing how the mechanisms underlying cognition relate to, and modulate, each other. Although dynamical models are often very abstract (do not map directly on to physical mechanisms) they should be linkable to these mechanisms in one way or another. More specifically, it should be possible to make sense of collective variables in terms of the operation of mechanical components. Thus Bechtel's (1998) claim that dynamical explanations only provide covering law explanations, in contrast to mechanistic explanations, misses the mark to some extent. I do not doubt Bechtel's claims that the heuristics of decomposition and localisation have been central to scientific breakthroughs in a variety of disciplines. I do, however, take issue with the implied claim that a mechanistic approach pushes us to think of the components of systems as representations that are computationally transformed.

---

[111] More specifically, we do not need to somehow provide a lawful relationship between some *a priori* objectively identified state of affairs in the world and a neural-bodily component state in order to make use of the idea that components may signal, influence, or modulate each other in ways that ultimately result in the production of adaptive behaviour. In other words we do not need to do all of the things that are usually thought to be necessary for constructing a theory of content.

The notions of *representation*, *computation* (when its formulation rests on the notion of representation), and even *information* are all fairly troublesome concepts as the earlier survey of theories of content pointed out. Dynamical analyses, by contrast, do not suffer from the sort of problems encountered by representational concepts, primarily because dynamical systems theory carries with it little in the way of theoretical baggage. As Beer (1995b, p. 129) notes: "[t]o say that something is a dynamical system is to say only that its future behavior depends upon its current state in some principled way, with no additional requirement that this state be interpretable as a representation or that the evolution of this state be interpretable as a computation." To argue that a system is representational, however, we need to assume that the states of elements of the system in question correspond to aspects of the world and that this correspondence is used to build subpersonal models and plans of action. As should be evident from the discussion in previous chapters, it is not clear how such a system might work. Indeed it is not even clear what a representation (or computation) *is* let alone whether people and other cognitive agents possess them. By contrast, the notions of *state variable*, *parameter*, and *dynamic* (rules of system state evolution) do not have such troublesome foundations. They are well understood mathematical tools. This is not to say that it is easy, or even necessarily possible, to use these tools to construct adequate models of cognitive activity. But if useful and reasonably accurate explanations of cognitive activity can be produced without recourse to representational or computational terminology, then there should be no principled objection to their use.

6. What Kind of Dynamical System is a Cognizer?

Living systems are cognitive systems, and living as a process is a process of cognition. This statement is valid for all organisms, with and without a nervous system.

Humberto Maturana & Francisco Varela (1980, p. 13).

Introduction

One has to squint quite hard to see anything particularly cognitive in Watt's centrifugal governor. Even Harvey, Husbands, and Cliff's evolved robots exhibit far simpler behaviour than creatures such as flies, cockroaches, and ants. So the dynamical analyses offered thus far will not convince the determined cognitivist to swap sides. Yet the point of rehearsing these dynamical examples is not to argue that dynamical systems theory (DST) has *already* provided a non-computational explanatory framework for cognitive activity but, rather, to suggest that it is worthwhile trying to find out whether dynamical concepts might provide a fertile medium for constructing an alternative explanatory approach. After all, as Beer (1995b) and others have argued, by itself dynamical systems theory is no more a theory of cognition than digital computational theory. Despite the fact that there do exist some helpful resonances between DST and the emerging interactionist programme, it is important that a dynamical approach is supplemented by a solid theory of the implementing substrate. The danger for the interactionist is that the lack of mechanical specification in DST will give the appearance that interactionism is a vaguely conceived explanatory approach. Therefore, in this chapter I want to bolster the interactionist themes hinted at by dynamical analyses and develop the outline of a non-cognitivist *mechanistic* underpinning for explaining interactionism. By *mechanistic* I simply mean a *theory of the implementing substrate* that can ultimately specify the sorts of parameters and variables that should be used in a full dynamical analysis.

First of all, I want to develop the idea that cognizer's are special kinds of physical dynamical systems that interact with their environments in *adaptive* – that is to say, system-preserving – ways. I will refer to such systems as *autonomous* (or *autopoietic*) systems. This view stands in contrast to the cognitivist point of departure which is to conceive of cognizers as information-processing systems (e.g., Vera & Simon, 1993; see Smithers, 1995 for a critical analysis). These two perspectives are, of course, not necessarily mutually exclusive – its perfectly possible to think of natural information-processing systems as deeply concerned with survival and bodily maintenance. However, I will argue that the differing primary foci of the two approaches steers the researcher in different directions both in terms of the kind of research focused on and the ways in which results of that research are understood.

This basic autonomy approach, then, serves to reorient one's understanding of the nervous system, the primary control structure of many cognizers. It seems that the autonomy focus supports a view of nervous system as a bodily coordinator and regulator rather than as a computer that processes information about the environment. In this view the nervous system is primarily in the business of modulating bodily processes and maintaining an adaptive structural coupling between the environment and the body.

This view of the nervous system can be captured in simulations and models that use connectionist neural networks. Although these systems have been traditionally used to model a parallel and distributed version of the 'in-the-head information-processor', recent research has showed how the strengths *and* weaknesses of connectionism are well-suited to the creation of embodied control structures for embedded agents. Embodied connectionism (Bechtel, 1997), artificial life neural networks (Parisi, 1997), and the Distributed Adaptive Control framework (Pfeifer & Verschure, 1992a, 1992b) reinforce the regulation and coordination view of the nervous system. Despite these promising demonstrations, it is still widely argued that connectionist architectures are poor for making sense of logical and context-independent cognitive activities. They are, in short, excellent for modelling *basic cognition* but not so good at making sense of *advanced cognition.* Yet connectionism's weaknesses are illuminating, because they are also commonly *human* weaknesses. The question is whether connectionism can be supplemented so that, at least weak, human advanced cognitive abilities can be simulated.

Autonomous Systems and the Dynamics of Being Alive

In the last chapter it was argued that dynamical systems theory provides us with some powerful and illuminating tools for grappling with new themes that fall out from ESD-related research, yet it is widely argued that dynamical systems theory does not provide us with concepts for distinguishing the truly cognitive from the non-cognitive. Some cognitivist-oriented researchers have used the tools of dynamical systems theory to make sense of several concepts central to traditional cognitive science. For instance, the notion of representation survives, albeit in a mutant form, within several dynamical analyses of cognitive activity (see, e.g., Elman, 1995). Thus it appears that it may be possible to supplement dynamical analyses with a *cognitivist* theory of the implementing substrate. Such a possibility will have major repercussions for an interactionist theory of cognition. If we are to avoid assimilating ESD-themes within a modified cognitivist framework, as interactionists believe we must do, then dynamical analyses need to be underpinned by a different understanding of the nature of cognizers than that existing in cognitivism. There exist several tantalising hints as to what this alternative underpinning might be like and, probably not coincidentally, these alternatives often make use of dynamical concepts. Here I make the broad claim that the alternative foundation we should utilise is what might be

called an *autonomy (or autopoietic) analysis of cognition* (Christensen & Hooker, 2000, in press; Maturana & Varela, 1980, 1988).

Briefly stated, an autonomous system is a special kind of physical system, a physical system that is *delicate* but *cohesive*. Its cohesion is essentially *active* and dynamic (nonstationary) consisting of a collection of ongoing and interdependent chemical reactions (the system's metabolism). The integrity of an autonomous system relies on the maintenance of these reaction dependencies. You cannot simply break a piece off an autonomous system and get two smaller autonomous systems. Instead one gets either one damaged, but still semi-functional system, and one dead system, or two dead systems. So autonomous systems are also *holistic* entities. Autonomous systems require raw energy (food) from beyond their boundaries in order to fuel their network of chemical reactions. They achieve this by actively seeking out raw materials and transforming them into the molecules required for metabolism. Autonomous systems are thus, of necessity, *directive* and *interactive* systems; systems that manage their relationships with their surroundings in order to get the needed resources for maintaining their own organisational integrity. Because most, if not all, natural environments are changeable and unpredictable, autonomous systems also need to be *flexible* in order to maintain their integrity. Flexibility can occur because autonomous systems exist on the boundary between stability and instability (thus their delicate nature) and can inhabit multiple ordered states in response to low energy interactions with their surroundings.

Roughly speaking, this autonomy approach holds that cognition should be understood within the context of the organisation of living, self-producing, cohesive, thermodynamic systems with directive organisations that channel needed resources into their structures. This view shifts the focus of our basic cognitive approach from a discontinuity approach where cognizers are viewed as simple physical systems (bodies) controlled by complex physical systems (neural information-processing computers) to a continuity approach which takes cognizers (body, nervous system and all) to be complex physical systems that can maintain their identities by shaping the interaction dynamics of the organism-environment system. Or, to put it more simply, the autonomy approach holds that cognition is strongly bound up with the job of being and staying alive, whereas the computational information-processing viewpoint focuses primarily upon internal transformations of symbols or representations (Christensen & Hooker, 2000, pp. 27-31)[112].

---

[112] These two perspectives are, of course, not mutually exclusive. I do not want to suggest that cognitivism's internal modeller/planner assumptions are not *compatible* with the idea that cognizer's architectures are primarily concerned with 'survival', only that they do not *begin* with these questions in mind. Evolutionary and adaptive questions are not central to most cognitivist research. Rather, these questions have been used by some to modify and elaborate already existing explanatory concepts and strategies.

The cognitivist perspective focuses our attention on the actual cogitative processes thought to be necessary for recognising objects, recalling past events, solving problems, and so on. The interactionist perspective, on the other hand, focuses our attention on how a particular subset of physical systems (animals) manage their interactions with their surroundings.

Cognitive Systems as Living Systems, Living Systems as Autonomous Systems

The Chilean neuroscientists Humberto Maturana and Francisco Varela (1980, 1988) are (in)famous for claiming that the terms *cognition* and *life* refer to the same sort of thing.

> *Living systems are cognitive systems, and living as a process is a process of cognition.* This statement is valid for all organisms, with and without a nervous system. (emphasis in original, Maturana & Varela, 1980, p. 13)

Under this characterisation all organisms, from paramecia and pine trees, to porpoises and people, are considered to be cognizers[113]. This idea may seem strange, but I believe that it nicely captures one of the fundamental features of cognitive creatures that cognitivism fails to highlight – that cognition is all about ensuring the *survival* of an organism[114]. Cognition is of a piece with life because organisms can only stay alive by responding appropriately to ('understanding') the environment in an adaptive manner that takes account of the dangers and benefits of various kinds of external events and objects. Of course, there exist huge differences in terms of the flexibility and scope of environment-coping abilities across

---

[113] I think that Maturana and Varela's claim has much to recommend it. I, therefore, resist the temptation by the likes of Mingers (1995) and Christensen and Hooker (2000, in press) to discard Maturana and Varela's understanding of cognition as life and preserve a more traditional notion of cognition being a property of animals (with nervous systems) or, even more specifically, of 'complex animals' (with complex nervous systems). My advocacy of the Santiago theorists' claim should not be understood as a claim that there are not very important differences in the behavioural strategies and underlying mechanisms of different kinds of organism. Nothing much really rides on where one draws the 'cognition/non-cognition line' as long as one's characterisation of 'the cognitive' accurately reflects the abilities and processes that exist in the relevant organisms. However, I think that Maturana and Varela's notion of cognition has the requisite 'shock value' to reinforce the important *continuity* in the adaptive mechanisms of all organisms from bacteria, to plants, insects, and higher animals. Cognitivism's information-processing approach has the unfortunate tendency to reinforce the view that psychologists and their other cognitive science colleagues can ignore or downplay the importance of understanding basic life processes in their studies.

[114] Here survival is meant in an evolutionary sense as the maintenance of systemic integrity of an organism until it can reproduce, where this condition holds for *at least for a critical number of members of a population.* The latter qualification is necessary because, as Millikan (1993) and others point out, evolutionary concerns relate to the ongoing reproduction of *populations,* not individual lineages. For instance, a group of bullfrogs may survive as a population even if *most* tadpoles never manage to reach maturity and reproduce. In some cases survival must continue past the point of reproduction because, as in humans, the survival of off-spring is contingent upon the protection and guidance of the parent.

different species. However, the point of the 'cognition equals life' claim is not to argue that humans cope with the environment in the same way that plants do, only that environmental coping is a central factor in defining a particular class of physical systems.

Living things are *autonomous systems* (Christensen & Hooker, 2000, in press; Varela et al., 1991). Autonomous systems are special kinds of physical systems that can maintain their systemic integrity (what we usually refer to as 'staying alive') over a wide range of environmental conditions. More specifically, autonomous systems are:

- *cohesive* systems (where that cohesion is active, flexible, and holistic);
- *self-generating* systems (they manufacture the components that make up their structure, rather than relying on their food to contain all of biochemicals necessary for metabolism);
- *directive* systems (they require environmental resources to provide the raw materials for self-generation and possess an organisation that channels those resources into their body structure. They are not passive systems that can rely on resources being available in their immediate surroundings);
- and *interactive* systems (they must modulate their structure in order to cope with environmental changes. They can move from dangerous situations to safe ones and change the local environment).

These features may seem a little opaque at this point, so it is worth looking at each in more detail. I will deal with the concepts of *cohesion* and *self-generation* in separate sections and the concepts of *directedness* and *interactivity* in the section *Interaction and the Use of the Environment.*

Cohesion

Living things are *cohesive* in an active, flexible, and holistic manner (Christensen & Hooker, in press, p. 4). A cohesive system, according to Christensen and Hooker (in press), is "one in which there are dynamical bonds amongst the elements of the system which individuate the system from its environment." (p. 3). Unlike gases, that have no internal cohesion, or rocks that are cohesive because of their passive and rigid (high-energy) bonds, the cohesion of living things such as cells is achieved through chemical bonds with shallow energy well interactions. These bonds are characterised by "short time scales relative to the life of the cell and must be continually actively remade with the assistance of external energy fluxes." (Christensen & Hooker, in press, p. 4). In other words, living systems are dynamic, ever-changing, whirlpools of energy and matter, that continuously take in resources, transform them into useable molecules, and expel what is not needed. The 'thing' that is the living system is not a static collection of material components but rather

an ongoing process or *organisation* (Maturana & Varela, 1980, 1988; see also Christensen & Hooker, in press, pp. 7-8).

Active Cohesion

It may seem that continually active chemical reactions that can give rise to macroscopic order, rather than just random turbulence, occur only rarely, if at all, in the physical world. Intuitively it seems that continuous activity should correlated with disorder rather than order. But this is not the case. Despite the fact that many of the physical systems that we are familiar with (e.g., high school chemistry experiments) exhibit randomness and disorder when active, there exist many thermodynamic systems that become *more* ordered when energy is pumped into them. In these systems the molecular products of the reaction do not thoroughly intermix. Instead like-molecules congregate in particular regions and this gives rise to ordered macroscopic patterns. Such systems are known as *self-organising* systems. To be self-organising a system must: 1) consist of a large number of components (e.g., lots of molecules), 2) have non-linear interactions between those components, 3) be *dissipative* in nature, and 4) be *far-from-thermal equilibrium* (Kelso, 1995, p. 16). A *dissipative system* is one that takes in ordered energy and matter in order to maintain or expand its own order, and expels disordered matter and heat (see Prigogine & Stengers, 1985). On the surface, these systems seem to defy the second law of thermodynamics (the fact that all systems tend toward maximum disorder or entropy). However, it has been argued that, while such systems increase the amount of *local* order, they actually increase the rate of overall universal disorder (e.g., Swenson & Turvey, 1992). Self-organising systems are *far-from equilibrium* (or non-equilibrium) systems because their ordered nature (organisation) relies upon there being continued energetic input into them. When this energy is taken away the system breaks down.

Perhaps the most famous example of such a system is the *Belousov-Zhabotinsky reaction* (see Prigogine & Stengers, 1985). When the chemicals in this reaction mix the system does not equilibrate into a uniform blend of molecules but instead self-organises itself into a brownish background with a foreground of moving blue concentric circles. The reaction exhibits two different kinds of ordered pattern: 1) a pattern of concentric circles that propagate outward, and 2) radially expanding pinwheels that cartwheel about a centre (Kauffman, 1995, pp. 53-53). This spontaneous order (describable in terms of a few basic collective variables) maintains itself for a while until the system begins to move to a low-order regime. The ordered phase can be maintained, however, by carefully pumping in the reactants at appropriate rates. This maintains the non-equilibrium nature of the reaction. This is the sense in which autonomous systems have an *active* cohesion. That is, their systemic integrity (order) relies upon an ongoing collection of low-energy interactions and chemical transformations.

Flexible Cohesion

Flexible cohesion amounts to the ability of a system to produce appropriate (coherent, ordered) activity over varying environmental conditions. Put simply, a flexibly cohesive system does not fall to bits when conditions change, but rather changes its structure to compensate for the changes. Non-equilibrium systems can exhibit this kind of flexibility because they can undergo phase transitions from one ordered state to another (or from one ordered state to a stochastic or random state) with changes in the values of control parameters (recall the discussion in the previous chapter). For instance, in the Belousov-Zhabotinsky reaction merely *shaking the petri dish* (changing a control parameter value) can push the system from the spiral pattern to the pinwheel pattern (Kauffman, 1995, p. 54). As noted in the discussion of dynamical systems theory, control parameters are often *nonspecific* in their contribution to order parameter dynamics. That is, "in no sense do they act as a code or a prescription for the emerging patterns ...[,]" (Kelso, 1995, p. 16) "... patterns form or change spontaneously with no specific ordering influence from the outside (and no homuncular motor program inside)." (Kelso, 1995, p. 58). This ability to move between different *ordered* regimes in response to environmental changes is what makes the cohesion of autonomous systems *flexible* (Christensen & Hooker, in press, p. 4). Cohesive *equilibrium* systems, like rocks, however, are not flexible in that they cannot easily change *states* and, when they do, they typically change from a cohesive unit (a rock) to a state of 'low cohesion' (a pile of dust or a pool of molten rock).

Holistic Cohesion

However equilibrium systems, like rocks, *can* be broken up without affecting cohesion properties of the rock[115]. In fact, the same is the case for simple self-organising systems such as the Belousov-Zhabotinsky reaction (one can empty half of the solution from a Belousov-Zhabotinsky experiment into another container and, provided reactants are pumped in appropriately, the same ordered patterns can be observed). However, this is *not* the case for autonomous systems because "the forces which binds [their] parts depend on globally organised interactions." (Christensen & Hooker, in press, p. 4) and serious disruption to one part of the system often results in the entire system failing. Thus, there must be more to autonomous systems than just the possession of self-organising properties.

---

[115] This is so because "[t]he bonds are localised in the sense that the strength of the forces which bind a molecule within the crystal lattice depend only on the connections with adjacent molecules. This localisation means that there are no essential constraints on where the boundaries of the rock must occur – if it is split the particularity of the rock's identity is disrupted, but the result is two smaller rocks with exactly the same type of cohesion properties as the original." (Christensen & Hooker, in press, p. 4).

Self-Generation

Autocatalysis and Metabolism

Self-organisation is not a sufficient criterion for defining life. Living systems seem to have an additional quality to that of self-organisation: they are in an important sense, *self-sustaining systems* or *self-generating systems*[116]. That is, living systems acquire food and turn it into the molecules needed for building and operating their bodies. Living things "must constantly seek out sources of ordered free energy with which to replenish dissipated cellular structures and sustain the capacity for the processes that acquire these resources (e.g. foodsearch) and repair damage (e.g. reconstituting damaged tissue) or avoid damage (e.g. escaping a predator)." (Christensen & Hooker, 2000, p. 9). They continuously rebuild themselves by creating useful chemicals from the raw materials that they acquire. This feature of organisms is at base what is meant by the term *metabolism* (see Boden [1996a] for a discussion of metabolism in this context). We humans, for instance, replace 98 per cent of the atoms in our bodies every year; we rebuild our stomach lining in five days, our skin in six weeks, and our liver in two months (Margulis & Sagan, 1995, p. 23). What is 'conserved' over time is not some particular set of materials (what Maturana and Varela [1980, 1988] call *structure*) but rather a kind of dynamic process or *organisation*.

Kauffman (1993, 1995) proposes that such self-sustaining systems are a rather predictable consequence of organic chemistry. He has shown theoretically that, given a large set of basic organic molecules (reactants and catalysts), and a set of simple relations between them, that it is almost inevitable that an *autocatalytic network* will emerge. Some reactions require certain reactants or catalysts to work and these reactants and catalysts are the products of other reactions that in turn rely upon other reactions for their existence. When the molecular components of every reaction are the product of some other reaction in the system, the circle of reactions closes and a network forms. Thus an autocatalytic network is simply a set of molecular reactions that work together in a cohesive, circularly organised, *closed* system[117]. Such systems exhibit holistic cohesion because they cannot simply be broken up into smaller, but fully functional, units. The entire network needs to function appropriately in order for the system to cohere. Kauffman (1995) argues that autocatalytic systems underlie life itself.

---

[116] Maturana and Varela coined the term *autopoiesis* (literally: self-building) to capture this idea.

[117] Of course autocatalytic systems require 'food' in the form of a steady supply of basic molecules to stoke up the web of reactions in the network once the network has formed. Importantly, however, these food molecules need only constitute a very small subset of those required for the many reactions in the network. They are the raw materials that are transformed in the various reactions.

At its heart a living organism is a system of chemicals that has the capacity to catalyze its own reproduction. Catalysts such as enzymes speed up chemical reactions that might otherwise occur, but only extremely slowly. What I call a collectively autocatalytic system is one in which the molecules speed up the very reactions by which they themselves are formed: A makes B; B makes C; C makes A again. ... Given a supply of food molecules, the network will be able to constantly re-create itself. Like the metabolic networks that inhabit every living cell, it will be alive. (pp. 49-50)

Autopoiesis and Boundary Producing Reactions

One problem with applying the theory of autocatalytic systems is that such systems, if they occur with real chemicals, would require that their component molecules be kept in relatively close contact with each other for the reactions to occur. Kauffman (1995, pp. 66-69) suggests that compartmentalisation may be necessary in living autocatalytic systems to stop the dilution of a system's molecules with non-system molecules so that a sustained and adequate concentration of reactants is available to both start and maintain an autocatalytic system. Kauffman suggests that compartmentalisation may occur by: 1) confining reactions to a surface (e.g., clay, a bilipid membrane) rather than a three dimensional volume, 2) by dehydrating the system so cleavage of long polymers by water molecules is reduced, or 3) using energy from exergonic reactions to drive the synthesis of polymers from molecules with high-energy bonds (e.g., food). In all these cases there must exist some form of boundary or container that holds the system together in a localised region.

This is where the notion of *autopoiesis* comes into its own. Maturana and Varela (1980, 1988) argue that an autopoietic system, like an autocatalytic system, is a self-producing and self-maintaining system. However their formulation differs in the claim that one kind of autocatalysing reaction within the system is a boundary producing process (set of reactions). Thus, autopoietic systems differ from autocatalytic ones in explicitly needing and producing their own boundaries. The prototypical example of a boundary producing process is the ongoing production of the cell wall by prokaryotic cells. Maturana and Varela thus defines an autopoietic system[118] as a

dynamic system that is defined as a composite unity as [sic] a network of productions of components that,

a) through their interactions recursively regenerate the network of productions that produced them, and

b) realize this network as a unity in the space in which they exist by constituting and specifying its boundaries as surfaces of cleavage from the background through their preferential interactions within the network .... (Maturana, 1980, p. 29 quoted in Mingers, 1995, p. 15)

---

[118] On autopoiesis see Maturana and Varela (1980, 1988), Capra (1998), Margulis and Sagan (1995), Mingers (1995), and Stewart (1995).

A living thing *qua* autopoietic system[119] is rather like a self-contained factory that continually rebuilds itself as its components wear out or are used up. Maturana and Varela argue that because autopoietic systems create their own boundaries, for maintaining their closed autocatalytic network, that they can be understood as *operationally* or *organisationally* independent from their surroundings. At the same time it is obvious that autopoietic systems are complexly coordinated with their environments. Maturana and Varela capture these opposing ideas by arguing that autopoietic systems are *operationally closed* systems that are *structurally coupled* to their surroundings.

Interaction and the Use of the Environment

Christensen and Hooker (in press) criticise Maturana and Varela's notion of autopoiesis on the grounds that it obscures the important interactionist fact that autonomous systems are not just coupled to their environments, but they are essentially environment-*using* systems. They argue that Maturana and Varela's emphasis on operational closure focuses our theoretical attention on purely internal organisational matters and ignores the many ways in which autonomy maintenance in organisms relies upon relations within the environment as well as within the organism[120]. They argue that "the principal issue for Maturana in respect of multicellular systems is whether they have the correct structure to manufacture all their own components within themselves, not how well they cope with their environment" and that this has led to a "pre-occupation with locating autopoietic closure [that] in itself contributes little or nothing to understanding the basic evolutionary processes driving the emergence of neurally complex, adaptable life forms." (Christensen & Hooker, in press, pp. 8-9). By contrast, in their autonomy framework "the paradigm is the system that actively, directively constructs and/or compensates for external dependencies and

---

[119] Although it is commonplace for writers to say that Maturana and Varela consider *all* living things to be autopoietic systems, Maturana and Varela are actually a little more coy. At times they seem to suggest that the cell is the only living entity that we can be sure is an autopoietic system and that the status of larger organisms is unknown. They suggest that it is possible that multicellular organisms may be *second-order autopoietic systems* (which is to say, 'operationally closed systems made up of a collection of structurally-coupled first-order autopoietic systems (i.e., cells)') rather than autopoietic systems *simpliciter*. For the purposes of this work I will assume that all living systems are autopoietic.

[120] I am not convinced that Maturana and Varela's approach is *inconsistent* with Christensen and Hooker's autonomy approach, but the point that the former focuses our attention on internal matters is well taken. The notions of operational closure and structural coupling are, as Keijzer (1997) notes, primarily directed at constructing a naturalistic framework that makes sense of the *nervous system* in a non-Cartesian manner. Issues of how the organism's organisation makes use of the environment are addressed rather vaguely in terms of structural coupling. Yet if the relations that hold in the environment provide fundamental constraints on what internal mechanisms need to accomplish, as the interactionist position claims (e.g., Rowlands, 1999), then the autopoietic framework will need to take on board the criticisms of Christensen and Hooker.

constantly changes itself as it manages its interactions to respond adaptively to its environment." (p. 8); or more simply, "the general problem for adaptive systems is to produce environmental feedback that supports autonomy." (Christensen & Hooker, 2000, p. 14).

To accommodate an increased focus on how autonomous systems *use* the environment to maintain their bodily integrity Christensen and Hooker introduce a number of useful terms and concepts. Christensen and Hooker claim that autonomous systems are physical systems that actively produce the conditions necessary for maintaining their own systemic integrity (autonomy). Self-organising systems such as the Belousov-Zhabotinsky reaction and Kauffman's autocatalytic networks will automatically disintegrate if they are not located in a resource-rich situation. Moreover, self-organising reactions do not particularly 'care' if the environmental conditions become such that the integrity of the system is threatened. It is true that many self-organising systems are able to exhibit orderly patterning across a variety of environmental conditions, but there is no sense in which these chemical systems attempt to *avoid* environmental dangers. Christensen and Hooker argue that autonomous systems exhibit a special kind of order across changes in environmental conditions. Simply put, autonomous systems produce 'spontaneous patterns' that place the system appropriately in the environment so that the spontaneous pattern forming is maintained.

Thus, organisms do not just sit about in the face of resource depletion or the introduction of dangerous environmental conditions. Instead they actively seek out needed resources and actively avoid threats. Christensen and Hooker (in press) claim that this capacity arises from autonomous systems' *directive organisation*. Systems with directive organisations use environmental resources and alter their position in the environment in order to maintain their autonomy. They do this, so Christensen and Hooker (in press) argue, by seeking out "energy gradients that can maintain their dissipative processes and also act to maintain, and sometimes modify and elaborate, the processes which enable the exploitation of such gradients." (p. 3). Energy gradients are useful relations that exist between environmental entities. Autonomous systems possess mechanisms that use these gradients to get to places that contain useful resources or are safe (among other things). So "[a]utonomous systems are *interactively self-generating*: they so interact with their environment and within themselves that they are able to acquire the needed resources and direct those resources into the reconstitution of themselves." (Christensen & Hooker, in press, p. 7, emphasis added).

Christensen and Hooker (2000) propose that we understand directive organisation in terms of the interplay of three major features of autonomous systems: the system's *norm matrix*, its *action generation processes*, and the agent-environment *interaction dynamics* (or *interaction process*). They start by embracing the, by now familiar, interactionist idea that behaviour is a joint product of agent-side constraints and environment-side constraints.

Rather than the agent containing detailed, inner instructions for every tiny movement in a behavioural sequence, agent-side mechanisms are understood to *perturb* the intrinsic dynamics of the environment "rather like the way inserting a stick into a fast-flowing stream modifies the pattern of the water flow." (Christensen & Hooker, 2000, p. 11; see also Keijzer, 1997, 1998a). Christensen and Hooker refer to the overall agent-environment dynamic as the *interaction dynamics* or *interaction process*. The organism's actions (or *action generation processes*) modulate the interaction dynamics so that the organism's *autonomy* is maintained. Autonomy is understood here as the global constraint (or 'goal state') of staying alive by avoiding tissue damage, repairing damage that does occur, and by continuously reconstituting the dynamic, self-organising network of reactions that is the organism.

Of course, this global goal involves coordinating many separate systemic requirements. An autonomous system, for instance, must balance its need for food with its needs to breathe, conserve energy, maintain body temperature, and avoid injury. Every organism has a set of idiosyncratic requirements that must be met in order for its autonomy to be maintained. Christensen and Hooker (in press) refer to these requirements as *closure conditions* or *normative constraints*. An organism's internal action-generation mechanisms must generate actions that perturb the environment so that these conditions are achieved. The organism knows whether or not, and often how well, closure conditions are being met because it possesses an evaluative *norm matrix* – a collection of *explicit norm signals*. These are internal signalling systems that monitor the state of each of the various autonomy closure conditions. For instance, the feeling of hunger is an explicit norm signal for nutritional needs. An explicit norm signal like hunger is best understood, not as an internally represented goal for an animal, but rather as a system whereby an organism's body sets up internal correlations between some internal events (e.g., low blood sugar levels) and other internal events (e.g., heightened neural and hormonal activity that support food acquisition behaviour). If closure conditions are not being met, or are close to not being met, the animal's bodily state will modify so that appropriate action generation processes are set in motion for modulating the interaction process in order to improve the chances of closure. For example, a cheetah that is getting hungry will begin to produce actions such as searching for prey, stalking, and hunting.

Making sure that all closure conditions are met is no simple process, for the various explicit norm signals that make up a norm matrix often conflict with each other. For instance, a hungry cheetah may be 'motivated' to chase its prey immediately but have to put this aside in order to fulfil the goal of avoiding injury from a nearby adult male gazelle. Thus, overall autonomy must be met within the web of tensions set up by a norm matrix.

Christensen and Hooker are aware that this broad picture of the organisation of autonomous systems must be supplemented by a set of concepts that enable us to

distinguish between simple adaptive behaviour and complex intelligent action. They do this by making a distinction between different kinds of *adaptive management strategies*. They wisely argue that there exist no hard and fast distinctions between the strategies utilised by different kinds of organisms. Rather organisms differ in the degree to which they are able to flexibly and anticipatively modulate the interaction dynamics according to the plasticity, generality, and complexity of their action generation processes and norm matrices.

The general trend Christensen and Hooker perceive in the move from low level adaptive management strategies to high level ones is of an increasing capacity for *self-directedness*. Self-directedness, as distinguished from simple directed organisation, derives from an ability to *modify* one's action generation processes to better modulate the interaction dynamics. Increasing self-directedness is important for animals that need resources that do not always have the same environmental signature. If that signature is relatively robust, then a simpler low order adaptive management strategy will do.

For instance, female mosquitos must feed on blood to acquire the necessary protein for egg-laying (Klowden, 1995 cited in Christensen & Hooker, 2000). They find blood courtesy of a CO_2 gradient tracking mechanism. Blood can be found because in their species-typical environments higher CO_2 levels are lawfully related to the presence of blood hosts. The mosquito, however, has no mechanism for searching for blood *as such* – that relationship is dealt with in terms of the external environmental regularity between the presence of CO_2 and blood hosts. If CO_2 gradients arose for reasons unrelated to animal respiration, or some animal managed to limit or hide its CO_2 production, female mosquitos would have no way of altering their behaviour for they have no control over their blood-finding mechanisms. "Thus, the mosquito's adaptive management strategy is low order because the information it uses to modulate its actions concerns only a very narrow aspect of the interaction process; most of the relations on which it depends are implicit." (Christensen & Hooker, 2000, p. 12).

Moreover, mosquitos cannot alter or tune their CO_2 gradient-tracking abilities in order to find blood more efficiently. However, even bumblebees exhibit an ability to learn to associate flower colours with relative quantities of nectar (Real, 1991), thereby systematically modifying their gradient-tracking abilities. Large-brained mammals, including humans, have even more sophisticated self-directed abilities. Increasingly complex self-directedness increases the *management horizon* of the organism. Christensen and Hooker use this term to refer to the proportion of the interaction dynamics that the animal can modulate with its actions. A mosquito only deals with a relatively small management horizon that is spatially, temporally, and functionally *local* in nature – *spatially local* because its tracking mechanisms can only respond to CO_2 concentration in the immediate vicinity of the mosquito, *temporally local* because the mosquito's behaviour

is only modulated by the current momentary CO_2 concentration and not past or likely future concentrations, and *functionally local* in that gradient-tracking only affects relatively quick, reactive changes to flight settings rather than providing information for modulating a broad range of behaviours. In more complex animals the management horizon is considerably broadened with increasing abilities to track complex kinds of gradients that incorporate *non-local* spatial, temporal, and functional parameters. For complex animals

> what is tracked is no longer a simple environmental gradient, like a CO_2 concentration gradient, but a combination of system interaction processes and environmental organization, as evaluated by system norms. Thus, cheetahs track something like "effective hunting", specified as a relationship between injury-free movement effort and prey character (kind, size, health, etc.), terrain style ecological location, as evaluated against injury risk, hunger, urgency, satiation, [and] potential ecological risk. (Christensen & Hooker, 2000, p. 17)

In an important sense, more complex animals can see and act more broadly (in terms of temporally-extended events), more deeply (they are potentially sensitive to more aspects of the interaction process[121]), and more flexibly (they can deploy a variety of responses, over longer periods, to detected environmental features) than creatures with lower order adaptive management strategies. These kinds of abilities are enabled, so Christensen and Hooker argue, by increased abilities to: 1) *anticipate* future events, 2) *evaluate* whether, and to what degree, their actions are successfully modulating the interaction dynamics, and 3) to *construct* more efficient gradient-tracking mechanisms (Christensen & Hooker, 2000; in press). They refer to the collection of all of these abilities as *self-directed anticipative learning*.

It is important to be clear that anticipation, evaluation, and tracking mechanism construction are not necessarily complex, conscious, effortful, or even 'representation-hungry' capacities, although they *can be* in mature humans. Anticipation merely involves the ability to deploy an action somewhat in advance of the sensing of a particular event and can be accomplished by such basic capacities as 'feedforward action', distal perception, low-level motor emulation, and simple conditioning, as well as high-level imagination and inference (Christensen & Hooker, in press, p. 11). Evaluation of action success is accomplished by explicit norm signals and can range from simple action-specific signals, such as proprioceptive stretch and pressure sensors, to relatively non-specific ones such as pain and hunger. The latter, claim Christensen and Hooker (in press), are "important for understanding learning because non-specific signals allow the system to modify its behaviour to better satisfy its constraints." (p. 13), whereas the former are tightly connected

---

[121] For instance, a mosquito would possess a deeper strategy if it could not only sense and use CO_2 concentrations, but also be able to sense the smell of blood. It could then perhaps coordinate its behaviour with respect to more aspects of the interaction process.

to particular actions. Constructive gradient tracking exists wherever an animal can improve its ability to access a resource via gradient tracking. This can occur via low-level tuning of neuromuscular systems (see Goldfield's, 1995, theory below) to high-powered reflective analysis that leads to pursuit of new cues and gradients (Christensen & Hooker, 2000, pp. 17-21).

Goldfield's Theory of the Ontogeny of Action Systems

Christensen and Hooker, building on the insights of Maturana, Varela, and others, provide a view of autonomous systems (cognitive systems) which takes animals to be an interesting subset of dynamic physical systems – systems that are dynamically cohesive, that require a constant influx of environmental resources in order to exist, and that produce systemic activity that ensures that those environmental resources are directed into the system. The general idea of autonomous systems can be embellished with insights from Goldfield's (1995) theory of the ontogeny of action systems.

Goldfield (1995) has constructed a theory of the production of action by animals that has many resonances with the autonomy theory outlined by Christensen and Hooker. Drawing on Kugler and Turvey's (1987) and Thelen and Smith's (1994) attempts to theorise cognition and action in dynamical systems terms, Reed's (1982, 1996) theory of action systems, and many other sources, Goldfield proposes that we should understand action development as depending on the way "in which perception-action cycles ... are used to assemble task-specific devices ..." (p. 165). He suggests that there are two major processes involved in action production: *assembly* and *tuning*. Assembly, as noted in earlier discussions of dynamical approaches to development, involves the self-organisation of different components of bodily subsystems into devices for accomplishing particular activities. These systems overlap anatomically, so that certain muscles or neural circuits may at one point in time contribute to a device for kicking and at another time a device for walking. These devices spontaneously appear as attractor patterns in a dynamical system whose variables and parameters derive from the constraints of various bodily subsystems as well as constraints of the structured surround and the demands of the task at hand. In other words, task-specific devices are like very complex versions of the different, macroscopic, low-dimensional states that arise in the Belousov-Zhabotinsky reaction. They naturally arise from the network of tensions and constraints between interacting variables and parameters. These constraints vary across multiple time-scales, across ontogeny as bodies develop and mature, across phases of perceptual learning as skills are gradually learned, and at the finest scale of actual performance as forces and influences are subtly modulated in response to perceptual feedback. These kinds of modulation lie at the heart of the concept of *tuning* which involves the "active exploration of underlying dynamics in order to modify parameter settings given task demands." (p. 166). People and many other animals actively explore the effects of varying the values of control parameters by

tinkering with levels of force applied to movements, angles, speeds, periodicity times, and the structures of environmental objects and layouts. So, Goldfield's notion of tuning relates to Christensen and Hooker's (2000, in press) claims that intelligent action requires an ability to learn how to improve interaction process modulation (perturb the interaction dynamics using action) by evaluating perceptual feedback about action consequences in light of explicit norm signals that signify whether, and how well, particular autonomy closure conditions are being met.

Summary of the Autonomy Perspective

Do these observations and claims get us any closer to formulating a statement about the theory of the underlying substrate that we should use in constructing dynamical models of cognizers? Goldfield (1995) provides a number of concrete examples of dynamical models that contain parameters that correspond to things like the mass of an infant and the stiffness and damping characteristics of a bouncing infant's spring (pp. 177-181). These are clearly much more specific applications of dynamical systems theory than those that are needed for a general theory of cognition. Christensen and Hooker (2000) are agnostic about whether dynamical models will ever be the best way of expressing our theoretical understanding of the structure of cognitive creatures. However, the autonomy approach does provide us with a framework that makes the following sorts of claims about the kinds of things (variables, parameters, relationships) that are important in an interactionist vision of the cognizer.

The central concept in this perspective is that of *interaction dynamics*. This involves a commitment to viewing behaviour as an interactively emergent property of coupled brain, body, and environment systems that distributes 'control' across all of the systems. Coupling is conceived of as the mutual perturbing or parameterisation of the intrinsic dynamics of systems by other systems.

The body systems are conceived of as *coherent self-organising systems*. They spontaneously assemble low-dimensional, ordered whole-body movements (which we can call actions). Actions perturb the flow of the interaction dynamics.

Organisms are theorised as *autonomous systems*. They have evolved because they have been able to maintain their autonomy (their dynamic systemic integrity) over variable environmental conditions that exist in their species-typical environments. Put another way, evolution has seen to it that the kinds of spontaneous activity patterns that are realised by organisms' bodies adaptively perturb the interaction dynamics. The maintenance of autonomy involves balancing a number of factors which can be understood as *closure conditions* or *normative constraints*. Thus, perception-action body systems are sensitive to environmental features. Indeed they should be conceived of as *gradient tracking systems*

which are systems that use the lawful regularities between environmental entities to lead the organism toward resources needed to satisfy autonomy closure conditions.

Perception-action body systems are understood to be *softly assembled task-specific devices* constructed from anatomically overlapping resources in neural, endocrine, immune, skeletal (link-segment), musculotendon, and circulatory subsystems. These task-specific devices can be *tuned* through experience. This requires mechanisms for evaluating whether and how well particular closure conditions are being met by the action-based modulation of the interaction dynamics. This function is realised by a network of internal perceptual devices known as *explicit norm signal systems* (the *norm matrix* being the collection of all of these systems).

In many organisms, especially in mammals, birds, and similarly complex creatures, this tuning is of a very flexible kind. *Self-directed anticipative learning* involves an ability to gradually increase the temporal horizon of environmental gradients so that *anticipation* is possible. By utilising experiential feedback about the consequences of action these kinds of animals can use their sensitivities to large slices of the interaction process in order to learn about and use the various environmental regularities that underpin gradients. This gives these animals a high-level of flexibility with regard to their behaviours they can produce in order to maintain autonomy closure conditions.

The autonomy approach then, focuses theorists' attention on the entire body of the cognizer as an environment modulation system rather than focusing just on the brain as the 'locus of cognition' as cognitivism is wont to do. However, it is important to understand just how the nervous system fits into the broader autonomy perspective. This is the issue addressed in the next section.

The Role of the Nervous System in Cognitive Activity

The Regulation/Coordination/Integration View of the Nervous System

The autonomy approach provides a backdrop for understanding the nervous system that is quite distinct from the information-processing perspective. Organisms are understood as systems that strive to maintain their autonomy. Their internal mechanisms are organised in a manner that makes this possible. In many organisms these internal mechanisms include a nervous system[122], a collection of specialised cells for setting up a distinctive kind of

---

[122] By 'nervous system' psychologists usually mean the central nervous system or, even more particularly, the brain. In what follows I will use 'nervous system' to denote the entire collection of neurons (that is, the autonomic nervous system and peripheral nervous system as well as the CNS) and their supporting cast of glia, astrocytes and so on. Recent research (e.g., Meller, Dykstra, Grzbycki, Murphey, & Gebhart, 1994 cited in Maier & Watkins, 1998) indicates that glial cells, which can outnumber neurons by a ratio of 10:1 in some animals, perform much more than a mere supporting role in neural dynamics. The boundaries of the nervous

internal signalling system that assists in the coordination and regulation of other cells in distant tissues and organs. However, as Maturana and Varela (1980, 1988) note, there exist many multicellular organisms that do not possess nervous systems and accomplish these coordinative and regulatory functions in other ways. Consequently, they argue that nervous systems are not necessary for cognition. Of course this does not mean that the kind of cognition that, say, plants are capable of is the same as that of organisms with nervous systems such as chimpanzees or bumblebees. In particular, nervous systems enable a fast, target-specific, rapidly replenishing, and potentially long-range chemical secretion system for bodily regulation, coordination, and integration. Reed (1996) argues that organisms that need to move quickly and need to move different body parts in a relatively independent manner, in order to maintain their autonomy (i.e., most animals with the exception of sponges and such like), require just this sort of coordinative system (see also Maturana & Varela, 1988). However it is important to realise that, even in animals, the operation of synaptic signalling systems is deeply interwoven with the functioning of other regulatory systems including other chemical secretion signalling systems (the endocrine system, the paracrine system[123]), the immune system, direct plasma membrane-bound signalling molecules that influence cells that are in direct physical contact, and gap junctions that directly join cytoplasms of interacting cells so that cells can exchange small-sized molecules (Alberts, Bray, Lewis, Raff, Roberts, & Watson, 1989). These systems work in conjunction to modulate an organism's bodily dynamics[124]. Neurons, then, are not the only 'smart cells' in the animal body. Indeed, although they have special features, neurons are more similar to other secretory cells than is often acknowledged (see Alberts et al., 1989, chap. 19). The picture that emerges from a close analysis of the role of the nervous system within animal bodies is that of a specialised subsystem of a whole bodily regulation, coordination, and integration structure that includes interactions between many different cell, tissue, and organ types. In sum, the role of the nervous system is to rapidly regulate,

---

system, especially as it is understood as the central 'body control system', are perhaps fuzzier than is commonly appreciated.

[123] In *endocrine signalling* specialised endocrine cells secrete hormones that flow through the bloodstream. In *paracrine signalling* local chemical mediators are secreted and affect the immediate environment of the cell (approximately 1mm) before being destroyed or taken up. Although the endocrine and paracrine systems are widely discussed, some sources cite as many as five different chemical signalling systems. For example, Wessells and Hopson (1988, chap. 37) include an *exocrine* mode (a hormonal system where substances are secreted into ducts and subsequently body cavities and surfaces rather than into the tissue space near the cell) and an *autocrine* mode (a sort of localized paracrine mode where substances actually only work on the cell itself).

[124] For instance, chemicals secreted from neurons can act in paracrine mode as well as synaptic mode (Alberts et al., 1989, p. 683).

coordinate, and integrate the different body regions of those multicellular creatures that we call animals. *Nervous systems do not control behaviour directly,* but they form a vital and important part of the behaving body that is the animal. In other words, the body as a whole is the 'controller' of activity (activity that perturbs and modulates the interaction dynamic in order to support the animal's autonomy) and the nervous system forms one vital part of this whole (see also Damasio, 1999, chap. 5).

Support for the Regulation, Coordination, Integration View

Clark (1997, p. 130) argues that, within the neurosciences, there exist several modelling ideas that fit in well with an interactionist approach to cognition and, what I am calling here, a regulation, coordination, integration view of the nervous system. My version of Clark's list includes:

1. An emphasis on the nervous system as a *sensory-motor coordinator*. This notion fits in well with Maturana and Varela's (1980, 1988) notion of the nervous system as a system for enabling complex *internal correlations* between activity in the sensory surfaces and activity in the motor surfaces. This stands in opposition to current trend within the cognitive neurosciences which seeks to map particular psychological functions on to the neuroanatomy.
2. An emphasis on the idea that the nervous system supplies *part* of the structure necessary for enabling cognitive activity (the equal partner notion). The activity of the nervous system is understood to perturb (and be perturbed by) the dynamics of other bodily systems and, via the body, the agent-environment interaction dynamics. Behaviour arises from the coupling of these different dynamic systems. Behavioural control and production is distributed across body, brain, and world.
3. An emphasis on the *distributed and decentralised nature* of the nervous system itself. Instead of viewing the organisation of the nervous system as a physical manifestation of the sense-model-plan-act schema, it is understood in a manner more in keeping with the behaviour layer subsumption architecture championed by Brooks (1991a, 1991b) and his colleagues. The brain is a mess of low bandwidth connections between sensory and action oriented mechanisms.

Sensorimotor Nature of Nervous System Control

Clark (1997) notes that "the biological mind is, first and foremost, an organ for controlling the biological body. Minds make motions, and they must make them fast – before the predator catches you, or before your prey gets away from you." (p. 1). Clark contrasts this view of the 'mind' (by which I presume he means the *brain*) with the popular vision of the mind as "a kind of logical reasoning device coupled with a store of explicit data – a kind of combination logic machine and filing cabinet." (p. 1). Whereas Clark's idea derives from

his analysis of research in philosophy of mind and artificial intelligence, a parallel argument has been advanced within the neurosciences by Elizabeth Whitcombe (1996). She notes that the dominant modern view of the brain (and more specifically the cortex) derives from a basic strategy of mapping higher psychological functions onto separate neuroanatomical regions. This strategy has a long history going back to Broca, Wernicke, and, in more recent times, Geschwind. In her paper *The anatomical foundations of cognition: Suggestions for a reinterpretation* she argues that this functional mapping approach is "a temporary expedient, a compilation of two sets of data: extrapolations from animal experiments, and the 'topographical pathology' of aphasia and associated disorders of higher cerebral function." (pp. 81-82). The assumptions that underpin this approach, she believes, are untenable, and the data are more consistent with a view of the brain/CNS as a system for coordinating various 'peripheral end-organs' so that behaviour is made possible. Whitcombe remarks that

> It may be useful to follow the lead given by the first experimentalists in central motor function, Fritsch, Hitzig, Goltz and others. Rather than abstracting the cerebral hemispheres, or the cortex, from their context in the nervous system and the body it regulates, may we not restore them to continuity not only with the brainstem and cord but also with the peripheral end-organs which determine the specific character of sensory and motor modalities. ... Function may be seen as the property not of some discrete focal area, or some fraction of connectivity in the cortex, brain or central nervous system; but rather as a property of anatomical systems, which consist of specialized peripheral end-organs, receptors and effectors, together with their *corresponding* neural organization, peripheral and central. (Whitcombe, 1996, pp. 84-85)

In effect, Whitcombe argues that the nervous system coordinates our sensory and motor surfaces so that basic life functions are effected; functions such as posture and locomotion, breathing, feeding, excretion, reproduction, and sensing. One can think of these sensory-motor circuits as equivalent to the basic behaviour layers in a Brooksian subsumption architecture. These basic functional subsystems are then modified and interconnected in various ways by evolutionary and maturational processes and through experience, training, and practice, so that they combine to realise what the famous neuroscientist Hughlings Jackson called, *broad* or *cultivated* sensory-motor activities such as ballet-dancing, bicycle riding, and, swimming. Whitcombe's main focus is to argue that our ability to *speak* likewise derives from a complex interconnection of basic human sensory-motor abilities for such things as respiration and feeding[125]. She contends that "[i]t could well be that

---

[125] This is likely to be viewed as a fairly controversial claim given the modern enthusiasm for modular and innatist theories of language mechanisms (e.g., Pinker, 1994). But Whitcombe's knowledge of neurobiology and language are impressive and her argument convincing. She is certainly not claiming that parts of the brain have not evolved to support language or that language can be acquired via some general learning mechanism. Rather she argues that there is no self-contained central language module (or set of modules) in the brain (which do all the important 'information processing' related to parsing, lexical access, and so on) and that sends signals to motor systems to output in one way or another (spoken, signed, written). Instead she

speech and spoken language has developed from routines devised to exploit the vocal and the articulatory possibilities of breathing and feeding arrangements peculiar to the human species, rather than arising mysteriously as some 'faculty' with which the human brain is uniquely endowed." (Whitcombe,1996, p. 86). So, Whitcombe claims that broad abilities, such as speech, are the *functional adaptations* of narrower sensory-motor systems and that they are "devised to exploit their properties." (p. 85). There is more than a faint similarity between these ideas and Goldfield's (1995) idea, described above, that task-specific devices are softly assembled from anatomically overlapping components.

A related claim is made by the neuroscientists Antonio and Hanna Damasio (Damasio, 1989, 1994, 1999; Damasio & Damasio, 1994) in their statement of their *convergence zone hypothesis*. The Damasio's argue that the brain is primarily a system for adaptively dealing with the environment's impact upon the body of the living animal. More specifically they argue that the world is 'interpreted' in terms of how it affects the body in a survival-relevant fashion. Damasio (1994) puts it this way:

> If ensuring survival of the body proper is what the brain first evolved for then, when minded brains appeared, they began by minding the body. And to ensure body survival as effectively as possible, nature, I suggest stumbled on a highly effective solution: *representing the outside world in terms of the modifications it causes in the body proper*, that is, representing the environment by modifying the primordial representations of the body proper whenever an interaction between organism and environment takes place. (Damasio, 1994, p. 230)

They arrive at this conclusion from considering a range of neurological disorders that suggest that the brain does not create detailed 'representations' of perceived or imagined entities, at higher-levels of the processing stream, from simple representational features that are registered in early processing mechanisms. Rather, like Whitcombe, they argue that the brain contains basic sensory and motor fragments in early cortices (that map 'knowledge' of the world) and nonmapped procedures (what Clark, 1997 calls *neural control structures*) for coordinating these fragments in later cortices.

The Convergence Zone Hypothesis: the Sensory-Motor and Decentralised Nature of the Nervous System

The impetus for the convergence zone hypothesis is the *binding problem*. This is the problem of how the 'content' of perceived and imagined objects, which is dealt with by anatomically separated cortices (of, for example, colour, shape, movement, and size), is bound together so that coherent activity can be coordinated. Conventional neuroscientific

---

argues that language, which is at root speech, is fundamentally grounded in the dynamic arrangement of various sensory systems (e.g., auditory system) and motor systems (e.g., oral-laryngeal systems) and the physiological and biophysical processes that support them (e.g., respiration systems, postural systems, head and neck movement systems).

wisdom has it that the binding problem is probably solved via the creation of more complete and integrated 'representations' of events in the more anterior multimodal cortices (Damasio, 1989, pp. 28-30). In other words, it is widely thought that early sensory cortices sort out the basic features of what is being sensed and then send on their conclusions to a higher theatre where the features are assembled into fuller pictures. One can see the influence of the sense-model-plan-act schema in such an assumption. However, such a 'processing cascade' model of the brain is at odds with the neuropsychological evidence. In particular, one implication of the processing cascade model is that damage to the later, integrative cortices should prevent a person from perceiving (or recalling) the various features of a perceived or recalled entity. Damasio (1989, pp. 30-31) notes that, under this traditional view, bilateral damage to these integrative cortices should: 1) "preclude the perception of reality as a coherent multimodal experience and reduce experience to disjointed, modality-specific tracks of sensory or motor processing", 2) "reduce the quality of even such modality-specific processing", and 3) "disable memory for any form of past integrated experience and interfere with all levels and types of memory". This, however, turns out not to be the case. Amongst other contrary findings Damasio and Damasio (1994) note that:

1. Damage to early visual cortices compromises the retrieval of features (e.g., color);
2. Damage to intermediately placed cortices leaves the retrieval of features intact, but may compromise the retrieval of knowledge pertaining to certain categories of concrete knowledge, that is, compromise retrieval of knowledge of some nonunique entities while sparing others;
3. Damage to the anterior-most cortices compromises retrieval of knowledge regarding virtually any unique entity or event (scene) but leaves intact retrieval of features, entity components, and nonunique entities. (p. 67)

For instance, a person with early cortical damage may not be able recognise or imagine the shape or colour of an acquaintance's face. They are effectively blind to one or more features of a viewed or imagined entity. Damage to intermediate cortices may lead to the inability to recognise or imagine the class or category of an entity even though they can easily describe its various features. Damage to more anterior cortices may reveal itself in the person's inability to recognise or imagine unique entities such as a particular person's face or a specific episode while preserving their ability to categorise the viewed entity and describe its features.

In light of these findings the Damasios propose a more decentralised and distributed view of the brain's role in cognitive activity. They suggest that the neural substrate that underlies cognitive activity is composed of three sets of systems: a set of feature fragment systems, a set of convergence zone systems, and a set of attentional systems.

The first set of systems, located in the early (primary and early association) sensory and motor cortices, serves as the basic 'knowledge base' for sensory and motoric knowledge.

Damasio (1999) refers to these systems as the brain's *image space*. This is the only place in the brain where explicit (mapped) content is found. In other words, all cognitive activity is represented neurally in terms of the neural features within this set of systems. Moreover, because of this, Damasio and Damasio claim that it is only activity within these systems that provides the content of consciousness (Damasio, 1989, p. 45). Recently Damasio (1999) has provided a much fuller analysis of the neuropsychological basis for consciousness. For Damasio core consciousness is enabled by second order mappings in the brain's dispositional space (see the next paragraph) that incorporate mappings of an object of perception (which may include the 'inner perception' of imagery and memory), the state of body (what Damasio calls the *protoself*), and the way in which the body is changed in response to the object[126]. Damasio argues that these second order mappings can then be experienced as 'images' courtesy of their effects on the 'image space' that consists of the early sensory and motor cortices.

The second set of systems that Damasio believes make up his decentralised view of the brain is found in the higher-order cortices (sensory and motor association cortices of different orders in the occipital, temporal, parietal, and frontal regions) as well as in some limbic structures (entorhinal cortex, hippocampus, amygdala, cingulate cortices) and the neostriatum/cerebellum. Damasio (1999) calls these systems the brain's *dispositional space*. Its role is to coordinate the time-locked activity in the image space made up of the explicitly mapped sensory and motor systems. That is, these systems consist of machinery that connects the various sensory and motor knowledge fragments in a coherent and adaptive manner, so that complex combinations of sensory information and motor activity can be coordinated[127]. These systems do not contain explicit 'representations' of knowledge but rather dispositional (nonmapped) neural patterns which coordinate and organise the feature fragment regions (Clark's neural control structures). Damasio and Damasio refer to these coordinative areas as *convergence regions* (of which there are hundreds) made up of smaller neural ensembles of *convergence zones* (of which there are thousands). Anatomically, convergence zones are neural groups "within which many feedforward/feedback loops make contact" (p. 71) that "direct the simultaneous activation

---

[126] Damasio (1999, chap. 8) suggests that these second order mappings may occur in various nuclei in the thalamus, the superior colliculi, and in the cingulate cortices.

[127] The 'fragment record' regions (the first set of systems) also exhibit a limited binding capacity "capable of binding features into entities" but they cannot "map non-local contextual complexity at the event level ..." (Damasio, 1989, p. 34). In other words, the binding capacity of the early cortices is not enough to promote adequate organisation of feature fragments for the support of such cognitive operations as understanding the category membership of an entity, recognizing the specific or autobiographical nature of an entity, or fitting the entity into a temporal sequence or episode (see also Damasio and Damasio, 1994, pp. 67-68).

of anatomically separate regions whose conjunction defines an entity." (p. 65). Convergence zones are reciprocally connected with more anterior convergence zones as well as more posterior zones and ultimately explicitly mapped feature fragment regions. In essence, convergence zones associate different kinds of sensory, motor, and somatosensory activity, and develop epigenetically as the organism experiences different combinations of worldly events, actions, and bodily activities (Damasio, 1989, p. 48). The associations that different zones realise work 'in reverse' when imagination or remembering occurs. That is, higher-level activity is fed back into more posterior regions so that images (explicit, mapped activity) are reconstructed based upon the animal's previous experience of co-occurring sensory, motor, and somatic activities.

The third set of systems that the Damasio's postulate "ensure the attentional enhancement required for the concerted operation of the other systems," (p. 70). Damasio and Damasio (1994) do not go into any detail about these systems and I am not sure that even they are very clear about them[128]. In any event, these systems seem to constitute another kind of neural control structure for coordinating mapped regions so that adaptive behaviour is produced.

Although the Damasios continue to use the language of representation and information-processing, their convergence zone framework is quite compatible with ESD ideas. Convergence zones can be conceived in non-representational terms as the schemas that coordinate perceptual instruments (Thomas, 1999) or the low bandwidth connections between augmented finite state machines in mobots (Brooks, 1991b). In both cases convergence zones can be understood as coordinating non-representational detection and action systems rather than world-modelling systems. Most importantly, their approach provides empirical and theoretical support for the idea that the CNS is primarily a coordinator of sensory and motor surfaces even in animals that possess complex behavioural and psychological abilities that cognitive scientists normally think of as far removed from sensory and motor concerns (both neuroanatomically and psychologically).

---

[128] Damasio (1989, p. 49; 1994, pp. 196-199) says a little more about this hypothesized set of 'attentional systems' but I am less than convinced that, as described, they add much to the basic framework. His basic premise is that an image-generating neural system will create numerous candidate images (solutions) for any reasoning problem and that basic attentional and working memory systems are required to make such activity tractable. A basic attention mechanism is required to single out a particular image to be utilised while a basic working memory system is needed, so he argues, in order to keep the image activated for a relatively long periods time. It seems to me that the whole notion is rather vague and impressionistic and conflates the experience of reasoning with what must be the case in the underlying substrate. In order to make more sense of this suggestion, it needs to be made clear why and how specific 'images' need to be singled out for attention and kept alive at a *neural level* of description.

The Partial Role played by the Nervous System

We have already seen that there exist a good number of researchers that claim that the operation of the nervous system cannot be understood except in the context of the action loops that it participates in with environmental structures and body structures (e.g., Keijzer, 1997). Indeed the brain, especially the mammal brain, seems to be specialised for using such loops. This may be why Kathleen Gibson (1996) suggests that "neuroanatomical plasticity in response to environmental input is so wide-spread and integral to the functioning of many mammalian brains that they could well be considered bio-environmental or biosocial organs." (p. 36). In what follows I will discuss three important senses in which the nervous system should be seen as a partial, albeit important, contributor to behaviour generation: 1) the conservation of neural processes across different behavioural forms, 2) cortical plasticity supporting functional invariance across morphological changes, and 3) the somatic marker hypothesis which holds that the body proper plays an important role in internal 'information processing'.

Sunfish, Salamanders, and Mobile Robots: The Conservation of Neural Processes

There exists considerable evidence that at least some neural structures and processes are conserved across related species and across ontogenetic changes but that these *same* processes contribute to *different* behaviours because non-neural bodily morphology differs. Specifically, there is evidence that neural motor patterns are sometimes conserved across morphological and behavioural changes in ontogeny and phylogeny. Lauder and his colleagues have discovered this sort of effect in different species of sunfish (*Centrarchidae*) and in salamanders (Lauder, 1992; Reed, 1996). In a study of the feeding behaviours of sunfish they found that the motor patterns, that is, the neural impulses that flow down the spine, are quite similar across twelve different species of sunfish despite the fact that feeding behaviour and jaw morphology vary enormously across the different species. The implication is that it is differences in morphology and other non-neural factors, rather than differences in neural activity, that underpin differences in behaviour. Lauder's research with tiger salamanders shows a similar phenomenon that occurs across ontogeny. Neural motor patterns are preserved across metamorphosis, but behaviour and bodily morphology exhibit marked changes. Interestingly, Smithers (1992) reports a similar conservation effect in his research with Lego situated robots. He has found that it is often easier to tune a robot's behavioural dynamics by altering the robot's body shape and structure rather than through reprogramming its control system. Indeed, he claims that body alterations are sometimes the *only* way to get the robot to behave appropriately.

This conservation phenomenon is consistent with the idea that the modulatory effect contributed by an animal cannot be understood only in terms of the unique states of the nervous system. Rather, action results from the nonlinear coupling of body systems and

neural systems. And often it is easier, or more likely, that behavioural changes will occur when non-neural bodily systems change against the background of a constant neural dynamics. Thus neural action patterns, including so-called central pattern generators, do not uniquely specify a set motor/behavioural pattern (Pearson, 1985 cited in Reed, 1996). Nervous systems are not *mechanically specific* systems (systems that specify a sequence of movements) but rather *functionally specific* systems (systems that balance bodily factors against environmental influences in order to produce an appropriate behavioural functional outcome) (Reed, 1996). And this stands to reason, for animals must produce relatively invariant *functional* behaviours (e.g., moving in a particular direction at a particular speed) in environments where terrain, windspeed, water currents and so on can vary enormously. Therefore, animals must evolve internal systems that can produce relatively constant effects across different environmental conditions. "[S]election will operate on a *functional system* that spans the central and peripheral nervous systems, and probably includes bodily morphology as well." (Reed, 1996, p. 72). In other words, the nervous system makes up part of a larger behaviour production system that is *functionally specific* in nature. Its role, as an inherently plastic system, is to regulate and coordinate the larger system so that bodily and environmental influences are adequately balanced and the appropriate interaction dynamics are realised. The nervous system's purpose is to "set up a kind of field of force that can be maintained as an invariant background against which further influences – perceptual, behavioral, biomechanical – play." (Reed, 1996, p. 73).

Monkeys Fingers and Temporary Deafferentation: Cortical Plasticity

The functionally-specific nature of the nervous system is nowhere better demonstrated than in experimental findings which demonstrate the plasticity of the cortex. The cortex changes rapidly in response to changes in what can be called *task-relevant morphology*. One can see this process at work in Merzenich and Kaas' famous research on the somatotopic cortical effects of altering the hand-finger morphology of adult owl monkeys (Allard, Clark, Jenkins, & Merzenich, 1991; Merzenich & Kaas, 1991; see also Thelen & Smith, 1994, pp. 136-140). They observed the neural effects of changing proprioceptive input to the brain by 1) amputating one or two of the monkey's fingers, 2) surgically joining and later separately two adjacent fingers, and 3) training monkeys to use a device for getting food that stimulated only the tips of one or two fingers (thereby changing their normal 'input' experiences). The boundaries of the relevant cortical maps corresponding to these morphological-functional systems were found to change over a fairly brief period. The maps associated with the amputated finger were colonised by the maps from adjacent intact fingers. The surgically joined fingers formed a single map from the maps of the previously unjoined fingers. Moreover, when the fingers were later surgically separated the independent finger maps reappeared. In the unusual stimulation experiment the monkeys

developed greatly enlarged maps for the fingers involved. Again these reduced in size when the training stopped.

Kelso (1995) reports similar research that suggests that such cortical reorganisation can be even more rapid and more radical than suggested in Merzenich and Kaas' studies. For instance, the somatotopic maps of limbs in monkeys that have had their limbs deafferented (by severing the nerves in the dorsal roots of the spinal cord that carry sensation signals to the brain) are completely colonised by normally distant *face* maps (Pons et al., 1991). Similarly, by using blood pressure cuffs to 'functionally deafferent' parts of the arm and hand of human participants experimenters can observe effects on cortical representations within 25 minutes.

All of this research shows that the nervous system, and more specifically the cortex, flexibly recoordinates and reintegrates the dynamics of non-neural body systems to ensure that a consistent behavioural function is maintained. The cortex "does not represent either the periphery of the body or the external environment with which the periphery comes into contact; *instead cortex is a vehicle for weaving behaviorally significant aspects of the environment into a system that has the capacity to preserve some varieties of behavioral units rather than others.*" (Reed, 1996, pp. 75-76, emphasis added). The cortex seems to serve as an important arena for establishing the task-relevant coordination of body part dynamics. It is something like a kind of massively complicated behavioural homeostat – a system for making sure that the right kinds of 'output' (behavioural functions) are produced by a system that is perturbed by changes to the body that may occur in a number of ways: through low-energy coupling with the environment (sensation and perception), high-energy coupling (injury, surgery, contact with environmental objects), and through natural changes to body (maturing systems, illness and disease, physiological changes, etc.). Clearly the notion of 'representational content' is not fully suited to a description of the nervous system depicted by these claims, for the CNS is portrayed as a balancer and modulator rather than a representer or information-storer[129].

Gut Feelings: The Somatic Marker Hypothesis

Antonio Damasio (1994, 1999; Damasio, Tranel, & Damasio, 1991) has been struck by the ways in which a number of neuropsychological disorders seem to relate to patients' inability to appropriately evaluate environmental events because of an inability to *feel* the appropriateness of certain evaluations or actions. Upon inspection of the evidence he

---

[129] This does not mean that determined representationalists will not and cannot produce representational versions of the regulation, coordination, integration view of the nervous system, only that the dominant picture has shifted from one of taking in information and transforming it to one of adequately balancing the multiple influences on the system so that its cohesion is maintained.

suggests that these patients possess damage to neural circuits that monitor and use the dynamical changes in non-neural body systems and this prevents them from using important *content* about the nature of experienced or imagined events. As was noted earlier, Damasio (1999) refers to the overall neural-mapping of bodily activity as the protoself. In his terminology non-neural body processes serve as *somatic markers* for the environmental interactions (in perception) and planning and reasoning (in imagination). Damasio combines these ideas with evolutionary considerations to claim that

(1) The human brain and the rest of the body constitute an indissociable organism, integrated by means of mutually interactive biochemical and neural regulatory circuits (including endocrine, immune, and autonomic neural components); (2) The organism interacts with the environment as an ensemble: the interaction is neither of the body alone nor of the brain alone; (3) The physiological operations that we call mind are derived from the structural and functional ensemble rather than from the brain alone: mental phenomena can be fully understood only in the context of an organism's interacting in an environment. That the environment is, in part, a product of the organism's activity itself, merely underscores the complexity of interactions we must take into account. (Damasio, 1994, pp. xvi-xvii)

So the perception (and imagination) of an event derives not just from the characteristics of the event-in-the-world, but also from the kinds of bodily processes (muscle states, heart rate, and so on) that are happening at the time. The nervous system is thus deeply embedded in the body proper – the body works with the nervous system to create a viable whole. On this view interactions with the environment occur via interaction with the body. In Damasio's (1994) words:

Primordial representations of the body proper in action would offer a spatial and temporal framework, a metric on which other representations could be grounded. The representation of what we now construct as a space with three dimensions would be engendered in the brain, on the basis of the body's anatomy and patterns of movement in the environment.

While there is an external reality, what we know of it would come through the agency of the body proper in action, via representations of its perturbations. We would never know how faithful our knowledge is to "absolute" reality. (p. 235)

Damasio argues that all of our cognitive activities necessarily involve input from non-neural body systems, that is, the internal milieu, the viscera, the vesitbular system, and the musculoskeletal system. The animal as a whole interacts with the environment. Events and situations encountered by an animal result in changes to the whole body state that includes neural patterns as well as changes in other organs, joints, muscles, tissues, and so on. Natural selection has seen to it that these bodily changes *anticipate* potential actions and bodily responses. So, for instance, the perception of a dangerous situation not only involves changes to our 'thoughts' and 'action plans', but also to changes in heart-rate (increased circulation of oxygen to facilitate flight or fight), perspiration (to facilitate traction), hormonal levels, and so on. These bodily changes are subsequently fedback to neural systems by the animal's internal perceptual apparatus and impart an important part

of the content or meaning of an environmental encounter[130]. Damasio argues that this bodily feedback provides the foundation for primary emotional feelings – innate feelings related to evolved survival-oriented evaluations such as the human snake-fear response. This bodily feedback monitoring can be understood to form part of the substrate underlying *explicit norm signals* for autonomy closure conditions (Christensen & Hooker, 2000, in press) or *value schemes* (Edelman, 1987, 1992). Humans and, no doubt many other animals, also have the facility to associate unique, personal experiences with particular bodily states and subsequently come to acquire often subtle and complex feelings about objects, people, and events for which there is no (strong) innate response. Damasio calls such acquired body-brain response loops *secondary emotional feelings*[131]. Christensen and Hooker (2000, in press) talk of self-directed autonomous systems modifying their explicit norm signals in the light of experience to better deploy interaction processes aimed at fulfilling various closure conditions. Secondary emotions appear to perform a similar role. Damasio's *somatic marker hypothesis* builds upon this framework. He suggests that anticipative bodily responses to events provide a rapid means of evaluating a situation. These 'gut feelings' massively reduce the possible range of responses animals have to a situation. In humans, merely thinking about a possible course of action can engage a body-brain loop which

> forces attention on the negative outcome to which a given action may lead, and functions as an automated alarm signal which says: Beware of danger ahead if you choose the option which leads to this outcome. The signal may lead you to reject, *immediately*, the negative course of action and thus make you choose among other alternatives. The automated signal protects you against future losses, without further ado, and then allows you to *choose from among fewer alternatives.* (Damasio, 1994, p. 173)

Damasio and his colleagues (Damasio, 1994, chap. 9) have found that people with damage to their ventromedial frontal cortex display an inability to exhibit socially appropriate (secondary) emotional responses to disturbing events (Damasio et al., 1991) and an inability to adequately evaluate and plan activities (Bechara, Damasio, Damasio, &

---

[130] Damasio (1994, pp. 131-134) suggests that the neural-bodily pathway for primary emotions begins with activity in the limbic system (especially the anterior cingulate and amygdala). This activity triggers changes in the hypothalamus, muscles in the face and limbs, the autonomic nervous system, neurotransmitter nuclei, and other unspecified brain internal responses. The hypothalamus activity gives rise to endocrine and other chemical activity in the bloodstream. The brain then receives feedback signals from the viscera, muscles, and joints.

[131] Damasio hypothesizes that secondary emotional feelings include ventromedial prefrontal activity (the substrate of *acquired* 'dispositional representations') with the machinery of primary emotions.

Anderson, 1994)[132]. Such observations support Damasio's contention that body-state feedback loops are central to even high-level, rational cognitive activity.

Damasio's (1999) most recent claim is that the activity of body-proper systems is also central to the phenomenon of consciousness. As noted earlier in the section discussing the convergence zone hypothesis, Damasio thinks of core consciousness as an outcome of second order mappings which incorporate neural maps of an object, the protoself (the state of one's body), and the effect of the object on the protoself. He speculates that emotions are bodywide, survival-relevant responses to an object or event and that feelings are first-order mappings of those bodily responses (feelings, rather than emotions, are probably equivalent to Christensen and Hooker's explicit norm signals). Damasio (1999) argues that having feelings provides an advantage over just having emotions because "[t]he simple process of feeling begins to give the organism incentive to heed the results of emoting (suffering begins with feelings, although it is enhanced by knowing, and the same can be said for joy)." (p. 284). Core consciousness then builds on this feeling process by creating a high order "feeling of knowing that we have feelings." (p. 285). Damasio suggests that such knowing is important for complex behaviour and sensitivity to the environment.

> [K]nowing is the stepping stone for the process of planning specific and nonstereotyped responses which can either complement an emotion or guarantee that the immediate gains brought by emotion can be maintained over time or both. In other words, "feeling" feelings extends the reach of emotions by facilitating the planning of novel and customized forms of adaptive response. ... In looking for a good reason for the endurance of consciousness in evolution, one might do worse than say that consciousness endured because organisms so endowed could "feel" their feelings. I am suggesting that the mechanisms which permit consciousness may have prevailed because it was useful for organisms to know of their emotions. And as consciousness prevailed as a biological trait, it became applicable not just to the emotions but to the many stimuli which brought them into action. Eventually consciousness became applicable to the entire range of possible sensory events. (Damasio, 1999, p. 285).

Thus, Damasio's understanding of the evolutionary function of consciousness seems to dovetail nicely with Christensen and Hooker's (2000, in press) idea that complex animals exhibit self-directed anticipative learning, which is, at its core, the ability to monitor one's own reactions to the environment and to use that monitoring constructively and creatively to adjust one's behaviour.

Maier and Watkins (1998) propose theory from the field of psychoneuroimmunology that is similar in many respects to Damasio's somatic marker hypothesis. Like Damasio they argue that non-neural body systems play an important role in mental life. They present a wealth of evidence for the idea that "a variety of behavioral, affective, and cognitive

---

[132] Primary emotional responses remain intact. Patients with damage to the amygdala region, however, exhibit dulled primary (and secondary) emotional responding (Damasio, 1994, chap. 7).

phenomena are driven by events in the immune system." (p. 83). The immune system serves as a kind of diffuse internal sense organ for recognising foreign substances within the body. When the system detects a foreign substance within the body it initiates a series of adaptive behavioural and physiological changes that they refer to as *sickness*. These changes include fever, increased slow-wave sleep, alterations in plasma-ions, shifts in protein synthesis by the liver, increased circulation of white blood cells, reduction in food and water intake, increased responsiveness to pain, reductions in activity, exploration, social interaction, aggression, and sexual behaviour, depressed mood, loss of attention, and interference with certain types of memory. Maier and Watkins argue that this natural sickness response can come to be associated with stressful situations (e.g., public speaking, an impending athletic contest) so that it becomes activated in those situations. They note that "The theory being proposed is that events in the environment categorized as stressors activate the same circuitry as is activated by infectious agents, they just enter the loop at a different location." (Maier & Watkins, 1998, p. 93). They go so far as to argue that immunological responses are evolutionarily more basic than fight-or-flight responses in animals and that the latter is largely built upon the substrate of the former. This accords with the general idea that the nervous system has evolved in the context of simpler, non-neural body systems and has become a specialised subsystem that works to modulate the non-neural systems rather than being a completely new and self-contained system for producing behaviour.

In sum, Damasio and Maier and Watkins suggest that evaluations of surroundings are fundamentally dependent upon the nervous system's couplings with body proper systems. The brain does not so much form *representations* of what the body is doing (although Damasio sometimes writes as if this is the case); rather actual 'processing' occurs in non-neural systems and the consequences of these non-neural dynamics serve as parameterisations of nervous system activity[133]. We can utilise Christensen and Hooker's

---

[133] Damasio hypothesises that it is possible that we do not always use bodily-feedback to inform our (high-level) cognitive activity. With experience a person's brain can develop 'neural short-cuts' that bypass bodily input (he calls them emotional 'as-if' loops). We, in a sense, can use our emotional/evaluative skills without having to activate complete bodily feeling-related processes. However he doubts that the feelings that arise from as-if loops "feel the same as the feelings freshly minted in a real body state." (p. 156). This could be because there are chemical as well as neural aspects to the mechanisms that underlie 'real' (body state-based) emotional feelings" (pp. 157-158) and because "the brain is not likely to predict how all the commands – neural and chemical, but especially the latter – will play-out and the resulting states depend on local biochemical contexts and on numerous variables within the body itself which are not fully represented neurally." (p. 158). Damasio, although still captured to some extent by aspects of the cognitivist vision, seems to want to argue that important adaptively significant processes occur in non-neural parts of the substrate underlying the production of behaviour.

(in press) distinction between low and high-order adaptive strategies to make deeper sense of this idea. In low order strategies, such as female mosquitos' CO_2 gradient tracking system, a large amount of the interaction dynamics occurs in the environment-side of the dynamic equation. For instance, the relationship between concentration of CO_2 and the presence of a bloodhost is entirely embedded in the structure of the environment. The internal gradient tracking mechanisms of the mosquito rely upon this relationship in order to contribute to the production of adaptive behaviour. The mosquito cannot modify its tracking mechanisms to deal with situations where the relationship is violated. Indeed, the mosquito has no way of 'knowing' about the importance of this relationship. The same idea can be applied to the interaction dynamics of the nervous system and systems in the body proper. Important relationships that exist purely within non-neural systems are not 'discriminated' or controllable (and thus modifiable) by the nervous system. Thus the nervous system is not sufficient for producing 'commands' for behaviour. It does not contain all of the requisite 'content' needed for behaviour production. Rather behaviour emerges from the coupling of body and neural systems.

A Distributed and Decentralised Locus of Cognition

All of the previously discussed theories and their supporting research demonstrate how the adaptive production of modulation processes is widely distributed throughout the nervous system and the body proper. The regulation, coordination, integration view of the nervous system encourages us to resist employing the sense-model-plan-act schema as a template for describing the neurocognitive operations. The convergence zone hypothesis most obviously challenges the idea that the brain instantiates some kind of processing cascade where simpler representations are accumulated and coordinated into more complex representations at higher processing stages. Schechtman (1997) argues that much of the evidence and argument deployed in favour of the idea that the brain is the locus of cognition is also consistent with a distributed view of the role of the nervous system and that "[t]he standard conception of the brain as the locus of mind stems ... from the imposition of a Cartesian conception of self on a materialist ontology." (p. 149). Schechtman provides arguments in favour of the ideas that: 1) damage to the non-neural body affects cognition, 2) that phantom limb phenomena do *not* show that full perceptual experience can exist without possessing a limb, and 3) that Penfield's (1958) experiments in stimulating different sectors of the cortex to produce memory-like experiences do *not* show that normal cognition is possible without bodily influences. The continuing tendency to want to think of the brain as the locus of the mind comes from the fact that the standard view "meets the ontological requirements of materialism with very little impact on the traditional conception of the subject. Aside from the issue of immateriality (which is admittedly a large one), everything else that is generally presupposed about the subject within dualism can stay more or less the same." (p. 160). However, Schechtman argues

that "questions about the legitimacy of the traditional view of the subject must be raised if we are to take our materialism seriously." (p. 161). Schechtman believes that our current knowledge of neural phenomena is more consistent with some kind of distributed view of the contribution of nervous system to cognitive activity.

An Interactionist View of the Nervous System

So what does the nervous system do in an interactivist-constructivist framework? I have argued that the best way to think about the role of the nervous system is as a regulator and coordinator of the body (although, even non-nervous system systems must also perform these duties to some degree) rather than an internal world-modelling device. The nervous system cooperates with the body to produce bodily change that affects the interaction dynamics. At a very simple level the nervous system is simply a sensory surface-motor surface coordinator. It facilitates the many adaptive internal correlations between these surfaces. 'Poke' a sensory surface with an environmental event and a (hopefully) appropriate bodily postural response will result. Appropriate bodily responses are ones that preserve the system's autonomy. One need not think of the innards as implementing computations over representations of things in the world. Rather, the nervous system is a self-organising system that reorganises in response to changes in control parameters (changes in internal or external sensory surfaces). The reorganisation in turn alters the dynamics of the body as a whole (postural change and movement through the environment).

Now, of course, this is too broad and vague a picture of the nervous system's function within the agent-environment dynamic to get us very far, but it serves as a guiding and focusing picture in much the same way as the Cartesian notion that 'inner pictures' of the world are created in the nervous system and used for thinking and the planning of action. In particular, the regulatory-coordinative picture helps us to think of the nervous system as providing important components for the assembly of perceptual instruments, the circuits that regulate the coordination of perceptual instruments, and systems for implementing performatory action.

Recall that perceptual instruments are anatomically overlapping neuromuscular devices[134] for detecting and measuring specific changes in the external world and within the body

---

[134] Note that perceptual instruments are not just constituted by neural components but also from structures in the body proper. A 'hefter' cannot be used to test for weight without the existence of the muscles within the arm, hand, and shoulder which react and deform in particular ways to the presence of moving masses due to their spring-like qualities, and the hand-arm skeletal (link-joint) system which has its own dynamics courtesy of its physical features (mass, oscillatory behaviour, momentum and so on). In an important sense these non-neural systems do not just enable the nervous system to interact with the world by providing a protective casing for the nervous system, they also modulate the nature of the 'input' to the nerves and provide a ready-

(Thomas, 1999; see also Goldfield, 1995, pp. 159-160). These instruments influence the activity of other instruments (i.e., they provide the substrate for epistemic action) and provide signals for initiating performatory motor activity (i.e., pragmatic action). Many perceptual instruments can be thought of as measuring aspects of things in the environment (such as the weight of a hefted object, or the relative location of a fixated object) or the effects of objects and events on 'environmental arrays' (e.g., the structure of the light, distributions of volatile chemicals in the air). Other instruments measure goings on in the body including the activities occurring in the sensory surfaces (e.g., muscles that move the eyes, the state of cochlear fluid in the auditory system [Dallos, 1992, 1997, cited in Thomas, 1999]). Christensen and Hooker's (1999, 2000) *explicit norm signals* can be fruitfully understood in terms of the testing of bodily states using inward-focused perceptual instruments. For instance, hunger (a norm signal relating to available energy levels) may be realised by a collection of instruments that test for blood sugar levels, distension of the stomach, and the like.

Multiple perceptual instruments must be coordinated so that coherent agent activity can take place. Thus, the nervous system can also be understood to contribute a 'perceptual instrument coordination system'. Much of the nervous system may be given over to *neural control structures*. These are "any neural circuits, structures, or processes whose primary role is to modulate the activity of other neural circuits, structures, or processes – that is to say, any items or processes whose role is to control the inner economy rather than to track external states of affairs or to directly control bodily activity." (Clark, 1997, p. 136). Thomas (1999) seems to be talking about a similar, although more specific, kind of thing when he discusses the neural *exploratory procedures* or *schemata* that "determine the sequence in which the various perceptual instruments are brought into play." (p. 222). These neural control structures can be understood to assemble individual perceptual instruments and the coordinated behaviour of multiple perceptual instruments (and performatory motor systems) in a task-relevant manner. In other words these neural control structures might perform the same sort of functions as the low bandwidth connections between the augmented finite state machines (AFSM) that make up behaviour layers and

---

made dynamics for the 'output'. Indeed, the neural part of perceptual instruments must modulate the natural non-neural dynamics found in such systems (see also Schechtman, 1997). Burton (1993) provides a detailed examination of the role played by non-neural systems such as vibrissae (whiskers), horns, quills, teeth, and tools (such as sticks) in haptic perception. In each case the non-neural system can be understood as 'transducing' action on its surface so that the nervous system interacts with a transformed stimulus pattern. For instance, our outer skin is composed of relatively stiff plates of keratin (called *corneocytes*) and it exerts a shearing force on underlying innervated tissue when touching an object. To put it succinctly, non-neural parts of the body are part of (most?) sensory systems and not merely supportive of them. They provide behaviourally important signal transformations on afferent and efferent neural activity.

connect lower layers to higher layers in Brooks' situated robots. Recall, that these connections within mobots are best viewed in terms of simple suppress, interrupt, and encouragement signals rather than as systems whereby detailed information is passed from AFSM to AFSM. It seems sensible to suggest that that neural control structures may also perform this low-bandwidth coordination role rather than a high-bandwidth representation building role. Within most of Brooks' robots these connections are hard-wired and thus not amenable to modification with experience. But real animal neural control structures are no doubt plastic in nature. Edelman (1987, 1992) has suggested a possible process for modifying connections between multiple 'categorisation systems' (e.g., neural maps that respond to specific qualities of a target such as colour, orientation, and movement). These maps develop via experiential selection[135] and are reentrantly connected by massively parallel and reciprocal connections. These connections are strengthened when activity is correlated in two or more maps and weakened when it is not.

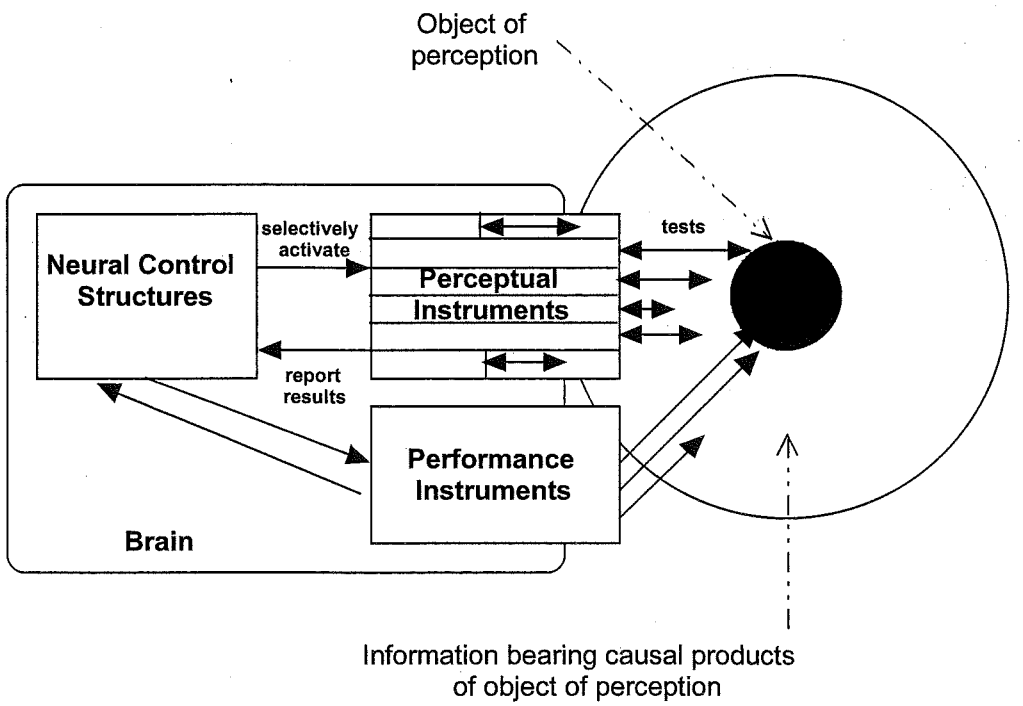

Figure 6.1 The Nervous System according to Perceptual Activity Theory
(modified from Thomas, 1999, p. 220)

---

[135] In experiential selection neural groups, or rather the connections that exist between neural groups, that work well according to some value system are maintained at the expense of less successful groups. Edelman (1992) notes that "during behavior, synaptic connections in the anatomy are selectively strengthened or weakened by specific biochemical processes. This mechanism effectively "carves out" a variety of functioning circuits (with strengthened synapses) from the anatomical network by selection." (pp. 83-85).

Action generation processes or *performance instruments* are those neuromuscular subsystems that accomplish particular kinds of pragmatic action. This idea derives from a bringing together of Christensen and Hooker's (2000, in press) ideas and Thomas' notion of perceptual instruments, but focuses not so much on systems for detecting information but for those that get things done.

As noted earlier, Goldfield (1995) discusses the idea that animals possess a *performatory action system* where various anatomical resources are softly assembled into task-specific devices[136]. For instance, the hand-arm system can play a role in devices such as a carrier, scooper, pusher, tapper, clapper, pounder, reacher, and grasper (pp. 260-261). Like perceptual instruments, performance instruments are made up of anatomically overlapping resources; not only overlapping with each other but, to a large degree, overlapping with the resources utilised by perceptual instruments.

The Autonomous Systems View in Summary

The hypotheses of the Damasios, Whitcombe, and Schechtman all point to a vision of the nervous system as primarily sensory-motor in nature. The brain is no longer viewed in a Cartesian materialist manner, as housing devices for carrying out different kinds of cognitive activity. Rather, the nervous system's function is viewed as the integration, coordination, and regulation of the body so that the changing world does not kill or seriously damage the animal. And frequently it accomplishes this in an *anticipative* manner by enabling the animal to be in the right place to get required resources or to avoid danger. The nervous system does not do all of this by itself however. It is essentially reliant on the states of other bodily systems and states of the environment in order to maintain autonomy. The nervous system itself 'merely' provides a fast, focused, and flexible matrix of internal correlations for perturbing the intrinsic dynamics of body and environment so that an appropriate interaction dynamics is maintained.

---

[136] Goldfield derives his ideas from Reed's (e.g., Reed, 1996) taxonomy of action systems. Reed suggests that it is useful to think of animals as possessing a collection of action systems in addition to set of perceptual systems (in Gibson's, 1966, 1979/1986 sense). The taxonomy includes a basic orienting system for maintaining posture, a locomotor system, an appetitive system for getting nourishment, a performatory system for moving and manipulating objects (in Reed, 1996 he calls it the manipulatory system), and an expressive system for modulating social interaction (In Reed, 1996; the expressive system is one of the interaction systems which also include the sexual reproduction system, the nurture and grooming system, and the semantic system). He also suggests that there exists a combination of the perception and action systems that he calls the play system. These systems all clearly relate to evolutionarily significant kinds of activity. However, my use the notion of performance instruments aims at a lower level of function than do Reed's action systems.

This view of the nervous system seems to be a more appropriate one for an autonomous systems approach than that given by cognitivism's computational perspective. Instead of theorising the agent as a machine that generates models of the world in its head and uses those to plan behavioural responses, the autonomy view of cognition takes the agent to be a flexible, sensitive, and complex system that adaptively modulates its activity with respect to the changing environment often with the express 'purpose' of altering that environment so that it can maintain its own autonomy. This is a complex view and one that requires the theorist to make some rather vague sounding claims about concepts such as *sensitivity* and *modulation*. The distinction between these terms and the theoretical terms of cognitivism can be better distinguished with a deeper understanding of dynamical systems terminology and a detailed knowledge of animal anatomy and physiology amongst other pre-requisites. However, these are generally not within the grasp of the average psychologist (including the present writer) and it thus becomes tempting to write off the claims of the autonomy approach and the regulation, coordination, integration view of the nervous system as a case of putting old wine in new bottles. One promising way of trying to make more concrete the differences between these new ideas and traditional cognitivist ones is to examine the ways in which connectionist models have been used to simulate the role of nervous system within an interactionist context. It is to these concerns that I turn in the next chapter.

7. Connectionism as a Model of the Embodied Nervous System

Mind is a leaky organ, forever escaping its "natural" confines and mingling shamelessly with body and with world. What kind of brain needs such external support, and how should we characterize its environmental interactions? What emerges, as we shall see, is a vision of the brain as a kind of associative engine, *and of its environmental interactions as an iterated series of simple pattern-completing computations.*

Andy Clark (1997, p. 53).

Introduction: Connectionism meets ESD Interactionism

I have argued that the autonomy approach combined with the regulation, coordination, integration view of the nervous system provides the basic framework on which an interactionist theory of the agent must be built. The salient question is: Can we be any more specific about the functioning of the nervous system within an interactionist perspective? In this chapter I want to suggest that the field of connectionism provides us with one of our best opportunities for understanding how the nervous system might contribute to the production of cognitive activity in conjunction with other bodily and environmental processes and structures. However, two problems immediately face the interactionist who wants to use connectionism.

The first is that, despite being a neurally inspired modelling approach, connectionism has been widely criticised for its biological implausibility. If connectionism does not reflect actual neural processes to some useful degree, its use within any theory of cognition opens that theory up to criticism (see Crick & Asanuma, 1986; Green, 1998; Reeke & Sporns, 1990; Segalowitz & Bernstein, 1997). I will not address this issue in any detail here; it suffices to say that the approach I take is to argue that connectionism, and its various incarnations in *biologically realistic networks* (Globus, 1992) and *synthetic neural modelling research* (Almassy, Edelman, & Spoorns, 1998; Edelman, 1987, 1992; Reeke & Sporns, 1990), is currently our best low-level take on how multiple interacting units can map inputs, understood as patterns of activity on sensory surfaces, onto sensible, adaptive outputs, understood as activity patterns on motor surfaces. I follow Farmer's (1990) proposal that the term *connectionism* should be used to refer to the general mathematical approach to modelling the properties of all kinds of dynamical systems where interactions depend upon a finite set of connections and where the connections are fluid. Such a characterisation includes neural networks, autocatalytic networks, classifier systems, and immune nets. We know that neural circuits, and indeed any array of cells, are networks in this sense. Thus we are at least looking in the right direction when we suppose that connectionist networks provide approximate models of the workings of the nervous system.

The second problem is that connectionism has largely been associated with attempts to model cognition as a purely in-the-head affair. In fact, at times connectionists have gone to great lengths to get their networks to mimic at least some of the internal operations of classical architectures. We see this, for instance, in Smolensky's (1991, 1995) attempts to model compositional operations by using the tensor product technique. Of course, it should be clear by now that interactionists do not believe that the whole cognitive story can be told purely in terms of the internal dynamics of a mind/brain regardless of whether it is modelled using a classical symbolic architecture or a connectionist one.

Despite these problems, I think that connectionism provides a promising framework for modelling the sorts of physical processes that may well go on in the workings of the perceptual instruments, neural control structures, and performance instruments that plausibly constitute a regulatory, coordinative, and integrative nervous system. My discussion takes the following form. First, I will briefly sketch out what connectionism is and indicate why so many cognitive scientists find it an attractive approach. Second, I will examine some of the problems that have been identified by critics. The third section consists of a description of two connectionist simulations which show connectionism's promise within an embodied and situated perspective. Finally I will examine several of the illuminating weaknesses of the connectionist framework.

The Connectionist Profile

Connectionist networks, also known as artificial neural networks (ANNs), are *nonlinear mapping systems* that are loosely inspired by the structures and processes of the nervous systems of humans and animals. Standard connectionist models[137] consist of several, usually three, layers of units, or 'neurons', that can map a coded input (coded in terms of the activation levels of the input nodes) into an appropriate coded output (again coded in terms of activation levels)[138]. A network is typically trained from an 'ignorant state', where the weight values between the various nodes are randomly assigned, by getting the network to construct outputs, based upon its current connection weight values, and giving it

[137] The 'white rat' of connectionism is the three layered, feedforward network that uses the backpropagation learning rule.

[138] There are certain limitations on these abilities. For instance, some pattern pairings require the existence of hidden layers between the input layers and output layers (e.g., the exclusive-or function, see Bechtel & Abrahamsen, 1991). As the complexity and variety of the input-output patterns increase, so too does the size of the network (number of units and connections).

feedback about the correctness of its attempts. This feedback is used to subtly adjust the network's weights using a learning rule such as the method known as *backpropagation*[139].

Varieties of Pattern Mapping

The mappings performed by ANNs can realise associations, transformations, 'recognitions', or completions depending on the kind of network used and the input-output pairings it is exposed to (Bechtel & Abrahamsen, 1991; Rowlands, 1999). *Pattern association* is simply the ability of the system to learn how to map an arbitrary input pattern onto an *arbitrary* output pattern. A variety of ANN, known as a *pattern associator*, can learn to pair, for example, an arbitrary code signifying the visual qualities of a number of stimuli (e.g., rose, steak) with a code that stands for the olfactory qualities of each stimulus (Hinton, McClelland, & Rumelhart, 1986). Some kinds of pattern associations can be thought of as *pattern transformations* – mappings that apply a particular operator to all inputs. For instance, Smolensky (1988) describes a network that can transform input data in accord with Ohm's Law. Many networks are also designed to categorise an input into a broader class. For instance, a network has been designed to map naval sonar returns (input) into one of two categories: mine present and mine absent (see Churchland, 1995). Connectionists refer to this kind of mapping as *pattern recognition*. Finally, connectionist networks called *autoassociators* can also engage in *pattern completion*. This is the ability to 'map' an incomplete input pattern onto a completed version of that pattern. In practice this involves activating a certain subset of units that represent a partial pattern and letting the network 'fill in the gaps'. In other words, networks can make good guesses about stimuli based upon incomplete information.

Typically the patterns that are mapped by simple networks are patterns that are extended in space but not in time. For instance, a word or an image representing a snapshot of a scene is mapped onto an appropriate output. We can call this *spatial pattern mapping* or *synchronic pattern mapping*. This sort of system works well if each input pattern corresponds only to one particular output pattern. For instance, it is possible to install a spatial pattern mapper in a robot that can provide appropriate motor commands to sensor input into a complex reactive mechanism akin to Brooks' subsumption architecture (e.g.,

---

[139] Backpropagation calculates the average difference between the actual output and desired output (the mean squared error or MSE) as a measure of the accuracy of the network and then, one weight at a time, checks to see if a slight change in the weight's value will reduce the MSE. Changes that result in a reduction of the MSE are implemented. This process is repeated many, usually thousands of, times and gradually the network (often) comes to 'learn' how to produce the appropriate outputs when given particular inputs (see Bechtel & Abrahamsen, 1991). There exist a number of other methods for training networks including other

Pfeifer & Verschure, 1992a, 1992b). However, many cognitive activities involve responding differentially to an input depending on where it falls in a temporally extended pattern. For instance, a mobot may need to produce one kind of motor output in response to an object as it approaches it and a different kind of behaviour as it moves away from it, even though, for a fraction of a second (a single input 'snapshot'), it may be receiving identical sensory input in both cases. The robot obviously needs some way to take into account how the particular input is embedded within a larger temporal context[140]. We can call this ability to be sensitive to temporal dependencies *temporal mapping* or *diachronic mapping*. Luckily for the connectionist, a type of ANN known as a *simple recurrent network* can do just this (Elman, 1990, 1995; see also Churchland, 1995). Simple recurrent networks use an extra set of context units as a kind of temporary memory of a previous input to produce networks that can respond appropriately to patterns that occur over time[141]. For instance, it is possible to train a simple recurrent network to recognise that a particular word (input pattern) may require a different output depending on the word's position within a sentence (e.g., whether the word is being used as a subject or an object).

Temporal mapping enables a network to engage in a limited kind of time-space distancing – that is, sensitivity to 'stimuli' that are distant in time. This amounts to being able to enter an event at a particular stage and, in a sense, have a good guess at what has gone before and what is likely to happen, and to modulate one's activity (output) accordingly. This

---

supervised training strategies, such as the *delta rule* (a close cousin of backpropagation), and unsupervised methods, such as *Hebbian learning*.

[140] In fact it is an entirely empirical question whether anything more than clever spatial pattern mapping is needed for this sort of robotic navigation. Pfeifer and Verschure's (1992a, 1992b) robot Lola, which is discussed below in more detail, makes do with a relatively simple ANN control structure. They note that "[o]riginally we had planned to get anticipatory behavior by introducing several network layers to keep some information about the recent past ..." but that "[t]his turned out not to be necessary." (Pfeifer & Verschure, 1992b, footnote 4, p. 26)

[141] Sometimes theorists talk of context nodes as being a kind of memory, perhaps a type of working memory, that holds in mind what was being thought at a previous time. I do not believe that this is a particularly useful analogy, for it does not actually make contact with the basic ideas in the working memory literature. For one thing, working memory research is typically concerned with the ability to keep in mind a series of items by rehearsing them. The role of context nodes in an recurrent network is to modify the impact of current input in the production of activity across hidden layer and, ultimately, output layer nodes. Although a loose analogy with some kind of memorial crosstalk may be read into the dynamics of simple recurrent networks, there is no strong theoretical motivation for thinking that simple recurrent networks deal in any kind of 'information-processing level' memory processes. Rather, I think that simple recurrent networks might provide some insight in to the patterns of activity that may occur in recurrently connected neural circuits. Such feedback systems may play important roles in amplifying or suppressing particular neural signals without thinking of them as a kind of very low level memory system.

gives networks a capacity for simple looking forward and looking backward in time that does not involve having to posit a detailed inner model of events (e.g., classically-inspired scripts).

Features of Connectionist Pattern Mapping

The structures and processes that underpin the pattern mapping abilities of ANNs are interesting for two reasons. The first is that the network stores its 'knowledge' of input-output associations across the entire network of units and connections, rather than being stored in a subcomponent of the system such as a register in a standard serial, von Neumann computer. So unlike traditional symbolic architectures, neural networks' 'knowledge' (i.e., their ability to appropriately pair input and output) is distributed over the entire, or large parts of the, network in the form of the networks' pattern of weightings between nodes. This feature is frequently referred to as *distributed representation* (Hinton et al., 1986).

The second interesting characteristic of ANNs is that they can store *multiple* 'representations' of input-output pairs within the network by using a single set of resources (connection weights). In other words, a single connection weight value can play a role in multiple input-output association representations. This feature of ANNs is known as *superposition* (Clark, 1993, pp. 17-23). Superposed representations have all sorts of fascinating properties that I will discuss in a moment.

The major advantages of such systems as models of human psychological abilities have been well-rehearsed in the literature (see, e.g., Bechtel, 1993a; Bechtel & Abrahamsen, 1991; Churchland, 1995; Clark, 1993; McClelland, Rumelhart, & The PDP Research Group, 1986). These unique selling points of artificial neural networks[142] include speed, robustness and graceful degradation, content-addressable memory, automatic generalisation, and automatic prototype extraction. Briefly, ANNs are *fast* (or at least potentially so) because they involve parallel processing. Most ANN research actually involves simulations run on conventional serial machines. Thus, this advantage is not often appreciated. Of course conventional symbolic models can also be implemented with parallel processing architectures, so this virtue is not *necessarily* distinctive of connectionist networks. *Graceful degradation* refers to the ability of neural networks to continue to produce reasonable output in the face of damage to the network. ANNs are much less brittle than conventional physical symbol systems. They also perform quite well when presented with incomplete or noisy data. They are, in short, tolerant of all sorts of abuse. *Content-addressable memory* refers to the ability of ANNs to automatically access

---

[142] At least those networks with distributed representations and nonlinear weight summation procedures.

appropriate 'stored information' without the need for a protracted serial search-and-match procedure. If one thinks of the output of a network as a retrieval of stored knowledge and the input as a cue or query for that knowledge, ANNs can be thought of as machines that can automatically 'look up' the appropriate memory without needing to search for it. ANNs can *automatically generalise* what they have learnt (the way specific input features and relations between those features relate to output features) to novel input. In a typical training and testing session a network is trained on a set of stimuli until it reaches a high-level of appropriate pattern matching. The network's weights are then frozen and it is tested on stimuli that it has been previously exposed to as well as a set of related, but novel, stimuli. More often than not the network responds appropriately to these new stimuli. Finally, Clark (1993, pp. 20-23) argues that one of ANNs' strengths is their ability to *automatically extract prototypes*. As is well known in the fields of categorisation research, a prototype is the statistical average of a group of similar stimuli[143], such as one's concept of the average dog that derives from one's experience of many types of dog (see, e.g., Medin, 1989; Rosch & Mervis, 1975; see also Churchland, 1995, pp. 49-53). Prototypes will not usually correspond to any actual member of a category. However many theorists argue that prototypes are used when we make typicality judgments and engage in other kinds of categorisation activities. Clark (1993) argues that ANNs automatically extract the statistical central tendency of feature complexes (prototypes of a set of inputs) because they incorporate mechanisms that strongly associate features via powerful mutually excitatory connections. The implication is that ANNs incorporate the human psychological ability for fuzzy categorising and typicality judgment 'for free'.

Some Problems with Traditional Connectionism

Despite traditional connectionism's intuitively attractive features, the framework has drawn criticism from a number of quarters. A number of cognitive scientists have argued that connectionism's holistic pattern mapping profile is inadequate for making sense of compositional cognitive abilities such as language, logical reasoning, complex problem-solving, and language-like thought (e.g., Fodor & Pylyshyn, 1988). Many connectionists disagree. I will leave discussion of these issues until the end of the chapter on the grounds that we should at least give connectionism a chance as a modelling tool for *non-compositional* cognitive abilities. If neural networks can be constructed to model such abilities this would be more than a minor success for the framework. However, even connectionism's ability to model these sorts of abilities has come into question and it is these critiques, I believe, that are fundamentally related to the issue of incorporating

[143] The notion of *similarity* is especially problematic for researchers of concepts and categories (Medin, 1989).

connectionism into an ESD interactionist framework. These critiques all derive from the problems of using connectionism within a conventional cognitivist information-processing framework. Traditional *simple connectionism*, to use Clancey's (1997) term, usually portrays ANNs as neural-like systems for implementing a version of cognitivism's mind as an in-the-head, world-modelling computer. Clancey (1997) argues, after Cliff (1991), that simple connectionism is severely restricted in its explanatory power when adopting the view of neural networks as pipeline processors that are isolated from real environments and thus require designed ontologies to make them work.

Pipeline Processing

Traditional connectionism follows the cognitivist assumption that the mind is an entirely in-the-head device. This is accomplished by thinking of neural networks as *pipeline processors* (Cliff, 1991, p. 32). Most current connectionist simulations use sequentially operating feedforward networks (even simple recurrent networks are basically feedforward systems) where the input layer is viewed as playing a sensing role, the hidden layer (usually there is only one) is thought of as modelling aspects of the world, and the output layer produces an 'action'. In other words, the dynamics of the networks are thought of in terms of a traditional information-processing sequence. Connectionist researchers often talk of the activity patterns of the hidden layer as realising internal representations stored in the network (although see Clark [1993] on the problematic nature of trying to understand networks in such a static and 'text-like' manner). In many ways this is quite an odd way of thinking of representation because the *input* layer activity seems to be a much more likely locus for representations of external objects, properties, or events than the hidden layer which is, in effect, a midpoint in the mapping process from a stimulus to a response. The concept of hidden-layer representations no doubt arises from the fact that statistics such as hierarchical cluster analysis turn up interesting kinds of partitionings (patterns of hidden layer unit usage) related to categorical similarities and differences among different inputs and the outputs that they map on to.

Sensory-Motor Isolation and Designed Ontologies

The second feature of traditional connectionism that inclines theorists to think of ANNs as incarnations of cognitivist in-the-head computers is its *sensory-motor isolation* (Cliff, 1991). Simple ANNs are akin to the philosophers' 'brain in a vat' in that they sit in splendid, unmoving, isolation responding only to input selected by an outside experimenter. In traditional connectionism it is the experimenter who supplies the 'perceived environment' (the input training set) and the consequences of acting in that environment (the set of correct mappings that are used to calculate output error which is used to change weight values). Clancey (1997) calls this experimenter-selected input a *designed ontology*. The neural network cannot choose its input by, for instance, moving its

sensory surfaces (input layers) using activity in its output layers in an independently structured environment. Thus, simple connectionism nullifies the possibility of perception-action loops playing a role in cognition and focuses attention on the ANN's internal dynamics as the locus of cognition. These features of simple connectionism fit it nicely into the traditional cognitivist approach.

The top-down approach to network training used in simple connectionism has been criticised as being derived from our folk-psychological and formal analytical assumptions about the appropriate kind of 'sense data' that is fed into cognitive systems (Hendriks-Jansen, 1996, pp. 75-81). Cliff et al. (1993) argue that only by exposing agents to real, low-level sensory input, as they do in their evolutionary robotic experiments, can we get an idea of how real agents utilise environmental information and, consequently, how real neural networks are structured (their connectivity and values of connection weights). Similarly, Hendriks-Jansen (1996, pp. 81-86) argues that we must carefully observe the behaviour patterns of animals, in the wild and experimentally, before we can make sensible suggestions about the kinds of stimuli to which they seem to be sensitive. Connectionist researchers, on the other hand, typically use highly artificial and abstract input sets (Clark, 1997; Cliff, 1991).

Cliff (1991) considers designed ontologies to be problematic because they provide an artificial scaffold for ANNs' supposedly powerful natural learning processes. This scaffold is thought to be problematic in the same way that toy worlds have been in AI research. The construction of toy worlds by researchers effectively *pre-processes* the input for the AI programs effectively doing much of the important and difficult work for the system. And because the pre-processing is carried out by the researcher, the question of whether such pre-processing could be accomplished by an extra subsystem bolted onto the central system is suppressed. (Often AI researchers have found that it is the pre-processing, typically perceptual and recognition, subsystems that are the hard ones to build, not the central systems that they concentrate on). This means that the internal processes of the 'intelligent system' might not, and typically do not, work in real-world environments because it proves impossible to implement adequate sensory and motor systems for coping with noisy, dynamic environments filled with novel objects and events[144] (see, e.g., Brooks, 1991a, 1991b). In effect this means that toy worlds can 'push' a particular vision of the workings of agents' internal systems (typically a cognitivist, information-processing view) at the expense of interesting, and possibly more workable, alternatives.

---

[144] In fact, the problems of how we can build adequate sensory and motor systems for traditional model-based architectures are central to the *frame problem* and Keijzer's (1997) *instruction problem*. See the discussion of these problems in chapter 3.

Verschure (1992; see also Pfeifer & Verschure, 1992a) has shown how this sort of problem occurs in Sejnowski and Rosenberg's (1987) NETtalk architecture. The purpose of NETtalk is to provide spoken (phonemic) output in response to written (graphemic) input strings. It turns out that the phonemic distinctions detected in the network's hidden layer activity occur because of the way the experimenters had constructed the input and output encoding systems and the set of correct letter-phoneme mappings in the training set. The units used to encode consonants differed almost entirely from those used to encode vowels. NETtalk then merely pulls out the patterns implicit in this encoding. The significance of the hidden layer partitions derives, not from the way the network uses its letter-phoneme correspondences (it does not use them at all), but from the constraints imposed top-down by the researchers. Clancey (1997) describes the problem this way:

> According to Verschure, the real work is in the creation of the 24-feature category system for pronunciation, plus the preparation of the 50,000 input-output pairs constituting correct behavior associations. In this respect, a parallel-distributed-processing (PDP) machine constitutes a clever hash-coding scheme: *The person outside represents an item* as a vector of features, which are numerically reconfigured for efficient storage. (p. 72, italics added)

Why is this problematic? Because, argue Cliff (1991) and others, a really useful application of connectionism can only occur if ANNs are allowed to deal with environmental input on their own terms. Most traditional connectionist research merely presents networks with *ungrounded* classical cognitivist symbols, both as input and as desired output, and asks the network to plot a mapping between the former and the latter. Connectionism, according to Cliff (1991), "has acted merely as a palliative for several of the maladies of symbolism." (p. 34). Just as past AI researchers could not find out much about the likely inner workings of real agents from simulations that use artificial micro-worlds, connectionist researchers cannot be expected to unearth potentially realistic network structures and dynamics by having them react to pre-processed input in the form of folk psychology-based symbols. Hendriks-Jansen (1996, p. 86) concludes that, while connectionism continues to rely on experimenter-supplied symbols for input (and as desired outputs) it will not have moved on much further from classical symbolism.

The Question of Representation

Central to many discussions in contemporary cognitive science is the question of whether or not connectionist systems are symbolic or representational in any interesting sense. By and large the majority of modern connectionists seem to believe that non-linear, distributed neural nets (see Hanson & Burr, 1990) are systems that trade in a special kind of non-symbolic representation (Smolensky, 1988). In classical symbol systems, symbols are tokens that stand for particular concepts. Typically they are thought of as being atomic entities that can be combined with other symbol tokens in a principled way according to

their "syntax" (roughly, their physical form) that maps directly on to their interpretation or semantics. Symbol tokens are thus thought to be *context-independent*; by this theorists mean that every use within a larger symbol structure (e.g., a proposition) of the concept, say, *coffee* uses the same token to represent coffee. Any contextual modification of the meaning of the proposition occurs in terms of the *external* syntactic relations holding between the coffee token and other symbol tokens (e.g., the tokens for **old** or **in a cup**).

Neural nets with superposed, distributed representations are, however, thought of as being *context-sensitive* (Clark, 1993). The idea here is that there does not exist any core physical token (say, an activation vector) that corresponds to a particular concept in all particular cases. Instead, there exist a number of related activation patterns that stand for such contextually-infected things as **coffee-contacting-porcelain** and **coffee-in-a-can**. There is no way, so the connectionists argue, of assigning particular meanings to individual units or unit groups that capture context-free features or microfeatures such that a single token can be found that underlies all uses of a certain concept. Instead, there exists a *family similarity* between activation vectors that deal with the same concept in different contexts. In other words, we can loosely speak of a variety of activation vectors as meaning a single concept, however their underlying differences will make a difference to further 'processing' within the network. So, for instance, my representation of coffee (in the can context) will quite possibly lead to different outputs compared to my coffee (in a cup context) representation.

Some critics of connectionism argue that this reveals a fatal flaw in neural network models of cognitive activity because only real, atomic, context-independent symbols can be used to generate the much vaunted quality of compositionality (systematicity and productivity) which underlies such cognitive abilities as language production and comprehension, and logic and reasoning (Fodor & Pylyshyn, 1988; Macdonald & Macdonald, 1995).

At any rate, most connectionists understand their networks as being representational in the sense that they store some kind of knowledge distributed throughout the network in the connection weights. However, there are an increasing number of researchers who do not think of neural networks as containing internal representations of environmental entities in the patterns of activity across hidden layers (e.g., Edelman, 1987, 1992; Hendriks-Jansen, 1996; Munsat, 1990; Pfeifer & Verschure, 1992a, 1992b; Wheeler, 1996).

There are a number of reasons for holding such a belief:

First, it is possible to claim that a network *knows how* to, or can enable an agent to, map certain features onto certain other features without thinking of the system as an environment-modelling device. Many cognitivists are of the opinion that mere 'stimulus-response' systems are not cognitive because they do not involve complex internal transformations (e.g., Fodor, 1986). Yet most connectionist networks are just complicated

varieties of stimulus-response machines. In fact neural nets are not good candidates for sense-model-plan-act systems despite the fact that connectionists often use three layer networks with an input layer, output layer, and a hidden 'central processing layer'. ANNs can be fashioned with two layers, a single fully interconnected layer (e.g., Boltzman machines), or more than three layers. Recurrent networks, especially artificially evolved, continuous time networks (e.g., Beer, 1995a, 1995b; Wheeler, 1996) can become close to impossible to model in terms of 'processing stages'.

The second reason for thinking of neural networks in non-representational terms comes from an appreciation of the methods connectionist researchers use to make representational statements about their models. The impression that ANNs' hidden layers contain representations has been reinforced through the use of a number of statistical techniques such as hierarchical cluster analysis, principal components analysis and contribution analysis (see Clark, 1993). These techniques summarise the ways in which the networks function when they are presented with a variety of test inputs. For instance, hierarchical cluster analysis gathers together pairs of similar hidden layer activation patterns; one can see from this which inputs are dealt with in similar ways. However, these analyses do not capture all of the subtle 'semantic nuances' of neural networks (Clark, 1993, p. 61). As a number of writers have claimed (e.g., Churchland, 1989; Smolensky, 1988), a complete understanding of how network patterns contribute to behaviour ('output') can only be had at the numerical level of weights and unit activation levels (Clark, 1993, p. 62). Clark (1993) argues that we need be careful not to think that the likes of cluster analysis provide a foolproof method for unearthing anything like traditional representations in the messy complexities of network dynamics. "The final call ... is for caution: Connectionism reconfigures content in dynamic ways which can often outstrip our attempts to capture it in a piece of static, text-like code." (Clark, 1993, p. 67).

The final reason for being sceptical about representational claims within connectionism arises from attempts to use networks within real robots. When networks are embedded in robots that move about the world and select their own input, it becomes more important to study the covariance of neural dynamics and *behaviour* than that between the neural dynamics and the objective environment (Pfeifer & Verschure, 1992a, 1992b). If anything, such analyses can be used in *retrospect* to work out what a robot is 'seeing' within its *Umwelt*. That is, one can legitimately hypothesise that if a robot behaves in a similar fashion in a variety of contexts that it 'sees' those contexts as somehow similar in kind. This is a completely different task from working out how a system builds internal representations of worldly objects from sensory data. In these cases representational analyses are just not as useful as alternative analyses (e.g., dynamical ones).

Embodying and Situating Connectionism

Not all connectionist simulations have been constructed as alternative, non-symbolic models of the cognitivist mind/brain. Pfeifer and Verschure (1992a) even argue that connectionism is *neutral* with regard to the broader explanatory framework adopted by the theorist. Indeed, many of the successful simulations in the connectionist literature involve analyses of skilful, sensorimotor activities (basic cognition) rather than compositional, high-level ones (advanced cognition). Networks have been used to explore such basic sorts of skills as face recognition, handwriting recognition (signatures, addresses on mail), visual processing, and motor control (see Clark, 1997). Bechtel and Abrahamsen (1991, chap. 5) and Rowlands (1999, chap. 7) argue that connectionism provides a fertile framework for understanding instances of, what Ryle (1949) referred to as *knowing how*. This is the kind of knowing involved in riding a bicycle, or playing tennis, or even carrying out experiments. It is skilful or procedural knowing[145] as opposed to *knowing that* or declarative knowing. Bechtel (1997; see also Bechtel & Abrahamsen, 1991) has even argued that our ability to engage in formal logic is better understood as an instance of *knowing how* to use external rules and symbols. He has even constructed a simplified connectionist network that learns, succeeds, and fails in formal logic exercises in much the same way as his undergraduate students do.

Yet, despite there being suggestive resonances in the literature, much connectionist research has failed to model cognitive activity in a realistic manner. Clark (1997) notes that connectionism's

> ability to illuminate biological cognition depends not just on using a processing style that is at least roughly reminiscent of real neural systems but also on deploying such resources in a biologically realistic manner. Highly artificial choices of input and output representations and poor choices of problem domains have, I believe, robbed the neural network revolution of some of its initial momentum. (p. 58)

He argues that much connectionist research has "leaned too heavily on a rather classical conception of the nature of the problems ..." (p. 58) by concentrating efforts on modelling isolated cognitive abilities, such as producing the past tense of verbs, often using arbitrary artificial codings of the relevant stimuli. Clark suggests connectionism would be better

---

[145] I use the terms *procedural knowledge* and *knowing how* interchangeably. It should be noted, however, that both Bechtel and Abrahamsen (1991) and Rowlands (1999) state that the notion of *procedural knowledge* is traditionally used by theorists who model *knowing how* as being realised by condition-action rules (e.g., Anderson, 2000). This, then, situates *knowing how* in the traditional physical symbol system paradigm and thus stands in opposition to connectionism. I follow Bechtel, Abrahamsen, Rowlands, and, indeed, Ryle, in believing that this gets things the wrong way around; *knowing how* is primary and *knowing that* is derivative. Or, to put it less opaquely, our high-level cognitive abilities are grounded in our basic sensory-motor skills which are underpinned by dynamical/connectionist-like principles rather than

served by embedding networks in realistic settings where real-world input is used in the production of realistic actions as, for example, Pfeifer & Verschure (1992a, 1992b) and Almassy et al. (1998) do.

What an ESD version of connectionism, as a model of the regulatory, coordinative, and integrative nervous system, requires is, firstly, the use of *real input* in a realistic manner and, secondly, a coupling with a body and an environment so that it plays a *partial role* in the generation of behaviour. Fortunately, there exists some connectionist research which shows promise in dealing with these issues. I will illustrate connectionism's potential as a model of these factors by mentioning a couple of suggestive simulations. The first example comes from the work of Pfeifer and Verschure (1992a, 1992b) with their situated robot Lola. Lola learns how to 'visually' navigate its real-world environment in real time based on feedback from collisions. The second example, from the work of Beer and Gallagher (1992; Beer, 1995a, 1995b), shows how a connectionist 'nervous system' serves as a controller for the movement of parts of an agent's body without necessarily 'containing' all of the knowledge necessary for enabling the movements.

Real Input: Pfeifer and Verschure's Distributed Adaptive Control Approach

The 'real input' criterion demands that:

1. The neural net be able to deal with continuous real-time input from the environment rather than merely ordinal sequences of input.
2. The input must originate from sensors that transduce changes in the gradients of some 'environmental array' rather than being abstracted, symbolic descriptions created by an experimenter (i.e., this is essentially Cliff's [1991] designed ontology objection).
3. Furthermore, this input needs to be a product of an agent's self-directed activity in its environment rather than a series of experimenter-controlled stimuli presentations.
4. And, consequently, learning and behaviour-production must occur simultaneously rather than in two distinct phases.

Pfeifer and Verschure (1992a, 1992b) have attempted to address these issues by constructing a mobile navigating robot that they call Lola. Lola is an example of an approach to using neural nets that they refer to as *Distributed Adaptive Control.* They argue that we will only produce useful behavioural models with neural networks if they are embodied (control some sort of 'body system') and operate in an environment where input

---

symbolic/computational ones. Lakoff (1987), Johnson (1987), and Glenberg (1997) have begun to try and elucidate this sort of idea.

is selected by the system itself (see also Parisi, 1997). More specifically, distributed adaptive control requires that the modeller use the following strategy (Pfeifer & Verschure, 1992a, 1992b; see also Hendriks-Jansen, 1996):

1. Characterise the robot's sensors and effectors and the nature of the environment in which it is going to operate (what objects inhabit it, how they are laid out, etc.). Lola possesses a collision detector, a target detector, a range finder, and a simple wheel-based motor system[146]. It was designed to work in an office-like, enclosed environment containing walls and a number of objects and apertures.

2. Define a *value scheme* for the robot. A value scheme is a set of basic hard-wired ('innate') sense-act reflexes that the robot possesses for achieving basic survival in its environment (rather like the most basic behaviour layer in Brooks' robots). Lola had three basic kinds of reflex in its value scheme: 'retract and turn in another direction after a collision', 'whenever a target is detected move toward it', and 'move straight ahead'.

3. Define the network architecture and the learning mechanisms that can implement the value scheme and provide a basis for modifying behaviour. Lola consists of a sensory input layer (range-finder input), an avoidance layer (the collision detector inputs to this layer), an approach layer (the target detector inputs to this layer), and a motor output layer (whose activity is determined by input from the previous layers and controls the steering and speed of the robot's motors). The layers are interconnected such that a modified Hebbian learning mechanism[147] can modify the weightings between the sensory layer and the approach and avoidance layers. This implements a kind of classical conditioning between the 'visual' readings of the rangefinder (conditioned stimuli) and the detection of collisions and targets (unconditioned stimuli) and the hard-wired unconditioned responses that constitute the value scheme.

4. Let the robot interact with its environment and analyse the resulting behaviour. Pfeifer and Verschure have analysed Lola operating in environments with and without targets. In all cases the robot developed the ability to rapidly visually *anticipate* collisions so that it navigated its environment without bumping into things. That is, the robot did not

---

[146] Lola began life as a simulation in a modelled realistic-physics environment. Thus the sensors and effectors, although modelled realistically, did not correspond to any specific kind of machinery. Lola's neural network was subsequently successfully exported to a wheeled robot that used unreliable infrared sensors. Strictly speaking only the real robot is called Lola.

[147] This mechanism incorporates an active decay function that enables 'learning' to occur continuously, but that leads to equilibrium behaviour in a stable environment. Learning need never be turned off in such a system (Pfeifer & Verschure, 1992b, pp. 26-27).

just learn to associate a rangefinder state with a collision state, but rather learnt rangefinder states that correlated with collision *avoidance*. It also learnt to back out of corners without hitting anything and follow walls (in a similar manner to Mataric's [1991] robot Toto) while searching for a target that was placed behind a hole in a wall. The robot learnt behaviours that were dependent upon the kind of environments it was placed in (context-relative learning) and on the structure of its own particular movement patterns. This meant that network patterns were not simply a function of what was in the environment, but of what the robot *did* in that environment (which was a function of its body structure, sensor and effector characteristics, and value-scheme reactions) over the course of its learning history.

In sum, with the distributed adaptive control approach, it is the robot itself (via its basic sensori-motor value scheme) that 'pairs up' input activations with output motor commands. The nature of the internal activation patterns cannot be anticipated by the researchers because they do not know how the robot will 'decide' to cope with its environment (or even if it will cope). Network activation spaces for such robots are personalised (relative to the machine's particular history of interactions – the robot in effect 'chooses' its input from one moment to the next), embodied (in the loose sense of being dependent on the nature of the sensors, actuators, and body structure of the mobot – a bigger robot will, for instance, have a larger 'personal space' than a smaller one and will thus need to react at different distances from objects to successfully avoid them), and situated (dependent on the kind of environment in which the robot learns). In fact, the weighting systems in even simple robots like Lola are likely to be idiosyncratic, that is, not 'decodable' in terms of their contribution to behaviour without a full understanding of the robot's history, body-structure, and environment structure. The sorts of analyses that are used in simple feedforward and recurrent networks (e.g., hierarchical cluster analysis, principal components analysis, and contribution analysis) will probably be inadequate for partitioning the hidden layers so that claims about internal representations can be made. One can analyse network dynamics in terms of how they correlate with activity (although this will change as learning proceeds), but this is a far cry from capturing some kind of permanent database of knowledge representations within the system which is used for recognising objects and planning action.

Thus Pfeifer and Verschure's distributed adaptive control simulations show how connectionist networks can be used to control an embodied and situated agent that utilises real input to produce adaptive behaviour. Their research also shows that the specific dynamics of a neural network, as a body regulator and coordinator, depend upon the nature of the robot's body (sensors, effectors, etc.). It, in effect, depends upon the 'processing' of environmental information given to it by the body-in-action. This observation is clearly

seen in the dynamical analyses performed by Beer and Gallagher on their connectionist leg controllers.

Partial Structure: Beer and Gallagher's Leg Movement Controllers

Although robot walking is probably not what most of us would think of as a cognitive activity, research into legged movement provides us with some nice examples of the ways bodies and environments cooperate with neural control systems to produce molar behaviour. Keijzer (1997, pp. 122-25) discusses work by the robotics researcher Raibert (1986, 1990; see also Hodgins & Raibert, 1991; Raibert & Hodgins, 1992, 1993 all cited in Keijzer, 1997) that gives a concrete illustration of these ideas. In particular he focuses on a one-legged robot (that, Keijzer calls 'Hop') that moves by hopping on a single compressible leg. Hop's leg is a bit like a shock-absorber in that it shortens when the robot impacts with the ground. This causes the air within the leg to compress and provide a means for propelling Hop upward in the next movement. The robot's control structure controls the direction in which the leg points when it hits the ground (via hydraulic actuators that control the position of the 'main body' with respect to the leg). This torso-relative-to-leg movement is used to balance the robot so it does not fall over and to tip the whole machine so that it moves in a certain direction. However, the control system "excites and modulates the bouncing motion, but does not specify the details of the trajectory ..." (Keijzer, 1997, p. 124) as occurs in classically designed robots. "Hop's control mechanism uses the dynamic interaction between the springy mechanical system and the control to generate motion" (pp. 124-125). Keijzer argues that Hop, and Raibert's other robots, provide support for the idea that body systems and neural control systems work as equal partners in the production of behaviour:

> Instead of thinking about a nervous system or a control system as a center for commands to be executed by actuators, the body and the environment are taken as a system with its own dynamic characteristics. The pre-existing dynamics of the body and the environment are only modulated by the controlling system, not really controlled. Raibert and Hodgins state: "We believe that the mechanical system has a mind of its own, governed by the physical structure and the laws of physics. Rather than issuing commands, the nervous system can only make 'suggestions' which are reconciled with the physics of the system and the task" (Raibert & Hodgins, 1993, p. 350). (Keijzer, 1997, p. 125)

It is just this sort of neural 'suggestion-making' can be seen at work in Beer and Gallagher's (1992; Beer, 1995a, 1995b) connectionist leg controllers. They have 'designed' a number of artificial neural network controllers for a simulated hexapedal robot. The impetus for this work was the creation of controllers for robot's such as Brooks' Genghis, Hannibal and Attila (Brooks, 1991b).

The robot body-design used by Beer and Gallagher consists of six rigid legs, each with a foot that can move up and down. Each leg is driven by three motor effectors, one

controlling clockwise torque, one anti-clockwise torque, and the other the state of the foot. When the foot is in the down position (presumably touching the ground) forces generated by the effectors move the robot (forward or backward depending on which way the leg is swinging). When the foot is up the motors swing the leg through the air (again forward or backward depending on which effector is active). Each leg is controlled by a small, fully-interconnected (continuous time, recurrent), neural network of five units. Three units are dedicated to driving the three effectors. The other two 'interneuron' units have no pre-assigned role. Each unit receives input from a sensor that measures the angle of the leg with respect to the robot's body. All six leg sub-networks are wired together to form a complete 30 unit movement network (see Beer, 1995a, pp. 186-189).

The goal of the simulations was to evolve a robot that moved forward in a robust manner (had an average forward velocity of greater than zero). This was, in effect, the 'selection pressure' that acted on the evolved networks. The weights, biases, and constants of the networks were artificially evolved using genetic algorithms (see Beer, 1995a, pp. 189-192 for more detail on the algorithms used and the gradual developments observed).

Eleven successful networks were evolved under three different sensory-input conditions: 1) input always available, 2) input never available, and 3) input available half of the time (the unreliable input condition). All of the successful networks exhibited the tripod gait typical of real, fast walking, six legged insects[148]. However the characteristics of the networks evolved in the different conditions differed in important ways.

Beer (1995b) refers to the networks that evolved in the 'always available' condition as *reflexive pattern generators* because they rely upon the sensory input about the state of the leg (its angle) in order to drive the dynamics of the leg network. When sensory feedback about leg angle was cut from these networks the robot would cease moving. Interestingly, because these networks were sensitive to leg angle changes, the simulated robot could adaptively modulate its behaviour to changes in terrain and changes in the morphology of the leg (e.g., shortening or lengthening of the leg). Evolution under the 'never available' condition resulted in the production of *central pattern generators* – networks that produced a toy-like, stereotypical movement pattern that was insensitive to changes in terrain and changes in leg morphology. When input was intermittent (the '50% available' condition) the networks produced generated a compromise solution where stereotyped movement was produced in the absence of sensory feedback and more robust and situation-sensitive movement was produced when input was available.

---

[148] Only a tripod gait will adequately support an insect in a balanced state. The creature's centre of mass must lie within the polygon formed by the supporting feet (see Beer, 1995a, p. 187, fig. 4).

The cases of the reflexive and mixed pattern generators are interesting because they provide illustrations of the ways in which the body can be understood to contribute an essential component of the movement control system. In an important sense these neural networks do not provide the entire mechanism necessary for the production of robust forward movement. These networks are akin to the partial, internal control structures found in Hop noted above. Beer and Gallagher's reflexive pattern generators rely on a control loop that includes the leg to complete the movement control structure. This is most clearly seen when one views the neural net activity from a dynamical systems perspective. Beer (1995a, 1995b) plots the dynamics of a single leg control structure in a three dimensional phase space. The values of each dimension correspond to the activation levels of each of the motor output units. He shows how one of the reflexive networks exhibits six different attractor dynamics dependent on the position (and direction of movement) of the leg (the control parameter). In effect these are six different phase portraits. They differ because of the changing influence of the control parameter. *In simple terms we can say that the network 'does not know what to do' unless the bodily dynamics to which it is coupled continuously changes in a particular orderly manner.* Without influence from the body, the leg movement will arrest at a particular point (a point attractor in the phase portrait) rather than exhibit the desired periodic trajectory (a limit cycle in the portrait).

The reflexive pattern generator examples are enlightening but it is likely, as Beer notes, that real animals (and robust robots) evolve in something like the noisy-sensor condition and thus possess something like the mixed pattern generators. This may seem like an opportunity to argue that, in real-life examples, the control system can be fruitfully thought of as at least partially unembodied and unsituated. On closer examination, however, it is clear that the significant behaviour produced by the network-leg system cannot be understood in this way. Recall that the mixed pattern generators work *better* when coupled with the body. One can think of this as the primary mode of operation – the 'no input' ability to produce stereotyped behaviour can be viewed as a kind of emergency backup (a bit like our ability to be able to briefly continue navigating through a room after the lights go out). Mixed pattern generators are, perhaps, simply failsafe reflexive pattern generators.

In sum then, what the leg controller example illustrates is that:

1. Body states can contribute 'input' to a control network such that when those body states become decoupled from the network the control system's performance degrades or even fails.

2. Thus the body (particular bodily states) in effect becomes part of the overall control system. The networks do not 'contain' the entire behaviour-generating mechanism. One could argue that failure in such a decoupling scenario is simply due to lack of adequate input for the network. In a sense this is correct. But the traditional cognitivist

vision is that input simply provides values for a fixed set of variables in internal programs. The command and control structure is largely fixed in place in the head waiting to be 'stimulated' (or have its parameters set) as theorists such as Chomsky (e.g., 1980) are fond of pointing out. And, according to this story, clever researchers should be able to unpack these internal control programs such that they can give us a good idea of the structure of the behavioural activities that they control.

The dynamical systems perspective, however, shows us that the 'input' that derives from bodily dynamics actually changes the whole nature and structure of the 'internal programs' and not just the values of variables (see Kelso [1995] on the idea that learning changes the dynamics of the entire learning system). On this reading it is probably going to be impossible to pin down an internal structure as the controller of the fine details of a behaviour[149].

3. It stands to reason, then, that evolution (in this case *artificial* evolution) selects for body-network couplings rather than independent, 'full-control' network architectures. Beer argues as much when he writes:

> In a dynamical approach to situated action, an agent's nervous system, its body and its environment are viewed as coupled dynamical systems. Given that bodies and nervous systems co-evolve with their environments, and only the behavior of complete animals is subjected to selection, the need for such a tightly coupled perspective is hardly surprising. Here the focus is on continuously engaging an environment with a body so as to stabilize appropriate coordinated patterns of behavior rather than the sequential sense-think-act processing cycle typical of computational approaches. (Beer, 2000, p. 11)

Minimally Cognitive Behaviour and the Scope of Connectionism within an Interactionist Approach

Examples of the use of connectionist networks within an embodied approach and in artificial life research (Parisi, 1997) are currently largely confined to modelling sensory-motor control systems and skills. Recently Beer and his colleagues have extended these ideas to the simulation of, what they call, *minimally cognitive behaviour* (Beer, 2000; Slocum, Downey, & Beer, 2000). They have artificially evolved continuous time recurrent network controllers for simple agents that can perceive whether or not an aperture is large enough to fit through, discriminate between their own 'body parts' and objects in the environment, successfully 'catch' a horizontally or diagonally falling object when only seen for a brief period at the beginning of its trajectory, and attend selectively to objects in

---

[149] Here is a toy example: Assume there exists an 'inner program' for determining greetings. The program has a single variable NAME which is filled by the input corresponding to the person you are talking to. The program is SAY "Hello, how are you NAME?". In a dynamical systems approach the 'input' might change the *whole program* from one dealing with greetings to one dealing with requests.

their environment so that two objects falling at different speeds can both be caught. In all these cases the agents are simulated and their environments are less than realistic. Nonetheless, they provide us with a glimpse of the ways in which a regulatory/coordinative nervous system may couple with body morphology and environment structure in the generation of adaptive behaviour. Although it is far too early to expect to have evidence for such systems being capable of modelling advanced cognitive abilities such as formal reasoning, it is worth asking whether it is likely that high-level cognition can be achieved by building bigger, more complex embodied networks? And it appears that there are good reasons why this is probably not going to be the case. It is to these issues that we now turn.

The Limits of Connectionism

Artificial neural networks are ideally suited to simulating skilful sensorimotor processes where long periods of training lead to gradual increases in the ability to coordinate behaviours and perceived events (Bechtel & Abrahamsen, 1991, chap. 5). This capability should not be underestimated, for it lies at the very foundation of many of the skills exhibited by most animals, including humans. ANN simulations nicely capture the fact that these *basic cognitive skills*, as I shall call them, are not just knee jerk type reflexes but subtly coordinated responses to aspects of the environment which are *discriminatory* (modulated by variations in the intensity and/or shape of the stimulus), *learnt and modulated by experience* (have a 'procedural memory' component), *deployed relative to the animal's state* (sickness, emotion, motivation, hunger, etc.), and potentially *anticipative.*

Thus connectionist ideas seem to get us some basic time-space distancing without having to endorse a representational (or at least symbolic) approach to cognition. It seems that embodied ANNs enable us to make contact with the past (via implicit procedural memory embodied in the network's ability to 'learn from experience'), look some way into the future via anticipating future stimulus patterns, and reach into the perceivable present (via sensory systems), and the nonperceivable present (via synchronic pattern completion). In sum, ANNs seem to give us a mechanism for at least modelling situated, context-sensitive cognition.

Moreover, the weaknesses of ANNs seem to incorporate features which are familiar to us when we learn procedural skills. Like us, ANNs are *slow learners*, requiring many practice trials to perfect our skills[150]. And like us ANNs can suffer from 'crosstalk' where "similar

---

[150] Although it is arguable whether people and other animals need quite as many learning trials as the average neural network does. Of course, most ANNs typically start their training from a state of complete 'ignorance' (randomly set initial weights) and have no internal guiding system (a value system or norm

encodings can interfere with one another (much as when we learn a new phone number similar to one we already know and immediately muddle them up, thus forgetting both)." (Clark, 1997, p. 60). The weaknesses of ANNs are also illuminating in at least one neurobiological way – they work best when they deal with small sets of domain-specific data. Too much variance in the kinds of input-output pairings limits ANNs' ability to learn. Instead ANNs seem to lend themselves to a kind of modular architecture where a collection of specialised systems are wired together by low-bandwidth connections (see Elman et al., 1996). And finally ANNs, like humans, seem to have real difficulty dealing with logical, compositional domains such as language (speech production and comprehension), off-line problem solving and reasoning, and 'displaced' creativity. Like humans, ANNs are good at frisbee but bad at logic (Clark, 1997).

However, although humans may be better at frisbee than logic, at least they can do logic when adequately trained. Neural nets, it seems, are fundamentally hamstrung when it comes to decomposing what they know (what they are able to do) into different bits of knowledge so that they can be recombined, substituted, or modified. Because neural networks are essentially holistic pattern mappers, they have no access to components of their knowledge and thus cannot engage in any cognitive activity which is fundamentally compositional in nature. This fundamental limitation of connectionism is discussed in two different, but related, ways:

The first problem with ANNs is that they do not seem to be able to embody an explicit syntax for ordering the knowledge components stored in the network. Because the knowledge stored in ANNs is context-sensitive rather than atomic in nature, they cannot distinguish the order or layout of stored knowledge components in such a way that they can be reordered or recombined in a logical manner that can support inference, systematicity, or productivity. Even simplified localised networks that have knowledge components mapped on to individual units face this problem, because the networks do not possess any principled (syntactic) method for ordering the knowledge components. Bechtel and Abrahamsen (1991) give an example based on the sorts of criticisms offered by Fodor and Pylyshyn (1988):

> [I]f **A&B** and **A** are two nodes in a network, the weight of the connection from **A&B** to **A** can be set such that activating **A&B** results in (causes) the activation of **A**. This could be viewed as a kind of inference, but the representation of *A* is not in any way part of the representation of *A&B*. Any two nodes could be wired to have the same pattern of influence, for example, **A&B** might excite node **Z**. Clearly, then, the connection is not compositional in nature, and the inference does not go through in virtue of the syntactic relation between the nodes. (Bechtel & Abrahamsen, 1991, p. 212)

---

matrix). Pfeifer and Verschure's (1992a, 1992b) show that by having a hard-wired value system learning can occur relatively quickly.

Efforts to augment ANNs by including a set of 'relation units' that assign roles to the different knowledge component structures do not seem to provide a satisfactory solution to the problem (for a connectionist at any rate), because ultimately such measures result in the connectionist constructing a neural network implementation of a classical symbol system (see Pinker's [1997, pp. 118-122] discussion of Hinton's [1981] proposition network).

Thus, those who argue that language and/or thought require a compositional (systematic and generative) architecture (e.g., Fodor & Pylyshyn, 1988; Pinker & Prince, 1988) have reason to believe that connectionism will not suffice as a description of the substrate of cognition. I shall not address these problems directly here, but instead direct the reader to Bechtel (1993b, 1996), Bechtel and Abrahamsen (1991), and Rowlands (1999) for thoughts on how neural networks may be able to model linguistic skill without succumbing to the charge of merely implementing a classical symbolic model.

The second problem derives from the *implicit* nature of the knowledge that is 'stored' within neural networks. Clark and Karmiloff-Smith (1993) argue that current 'first-order connectionist systems' are nice candidates for explaining the basic implicit levels of cognitive functioning found in humans, but they cannot account for the sorts of higher cognitive activity exhibited by people who seem to use more explicit 'representations' of knowledge – that is, knowledge that can be prised out of the internal milieu and to some extent examined by the neural net/person and used in different ways. Among the various skills endowed by being able to extract 'knowledge components' Clark and Karmiloff-Smith include: 1) example-independent knowledge of rules, 2) coping with systematic changes to the structure of the rules governing a domain, 3) intelligent self-debugging without full scale retraining, and 4) integration of activities with those of other subsystems operating on data included in different formats. Let's look at each of these skills in turn.

Example-independent Knowledge of Rules

Neural networks cannot, and do not, construct an activity pattern that actually represents a generalisation or rule. Rather, such 'rules' emerge from the total dynamics of the network. There exists no way to extract such a rule so that it can be explicitly used. That is, unlike neural networks, in people "[t]he generalisation, insofar as it would constitute information for the system, constitutes a piece of knowledge over and above the ability to judge each individual case correctly." (Clark & Karmiloff-Smith, 1993, p. 494).

Coping with Systematic Changes to the Structure of the Rules Governing a Domain

Thus, we cannot isolate *part* of a rule system realised by a neural net and replace it with a new component so that a new rule is implemented. For instance, a neural network that produces pattern transformations that respect Ohm's law (V=CxR) cannot simply replace part of its knowledge so that it can implement V=C+R (Clark & Karmiloff-Smith, 1993,

pp. 491-492). Only a complete retraining of the net from scratch can bring about such a change.

Intelligent Self-debugging without Full Scale Retraining

Similarly, a network cannot improve its performance by isolating a problematic knowledge component so that it can be retrained in isolation from the rest of the adequately functioning network. Clark and Karmiloff-Smith (1993) illustrate the ability to self-debug with a golfing example: "a golfer whose game suddenly deteriorates may isolate the cause of the trouble as the wrist component of the swing. This enables her to take focused action; instead of learning the whole swing anew, she can keep most of it intact and concentrate on the source of the trouble. Such a focus is not possible for a first-order connectionist system in which all its knowledge is inextricably intertwined and no control structures exist of capable of unpicking parts of the web while preserving others." (Clark & Karmiloff-Smith, 1993, p. 492).

Integration of Activities with those of Other Subsystems Operating on Data Included in Different Formats

The intuition Clark and Karmiloff-Smith express here is that one cannot transfer 'knowledge' that is distributed throughout an entire, perhaps idiosyncratic, network to another network (as may be the case for knowledge transfer between modules or people). What is needed is a symbolic summary of the knowledge that can be transmitted and that can then be used to reformat the receiving network so that it can internalise that knowledge.

They thus conclude that because currently existing connectionist systems are 1) example driven (they exhibit no rule-based or 'insightful' knowledge changes), 2) possess purely emergent knowledge (rules do not depend upon symbolic expressions that stand for elements of the rule), and 3) they have no self-generated means of analysing their own activity, that they are too inflexible to account for genuine thought. As they say, "[n]o system in which rules are always merely implicit and emergent can, in our view, exhibit the kinds of higher-order flexibility and creativity found in humans. Only explicit rules have the genuine, systematically manipulable components that make radical flexibility possible." (p. 504)

They put this claim to the test by rehearsing findings from Karmiloff-Smith's developmental research (Karmiloff-Smith, 1992, 1994). Karmiloff-Smith and her colleagues have conducted a good deal of research that analyses trends in the development of a wide range of cognitive abilities including aspects of language and drawing. She

perceives a common pattern in these developmental sequences[151]. Early on children's cognitive skills are 'data-driven' by the external environment and realised by holistic procedural knowledge. They appear to have an implicit understanding of the task and cannot rearrange or utilise components of their knowledge. Such an inability mirrors the limitations of first-order connectionist systems. As children (but significantly, not other animals [Karmiloff-Smith, 1994, p. 694]) achieve a high level of competence at this implicit level (Level I), they begin to transcend their knowledge-use limitations[152]. Karmiloff-Smith argues that there has been a reorganisation of the cognitive system and the implicit representations of knowledge have been *redescribed* in a more explicit format (Level E1). Karmiloff-Smith suggests that there exists a variety of levels of explicitness of representational redescription. Level E1 is revealed in the componential nature of the behaviour itself though the child is not aware of, or able to, express the ability. In her words:

> The redescriptions are abstractions and, unlike level-I representations, they are not bracketed (that is, the component parts are now open to potential intradomain and interdomain representational links). The E1 representations are reduced descriptions that lose many of the details of the procedurally encoded information. ... Once knowledge previously embedded in procedures is explicitly defined, the potential relationships between procedural components can then be marked and represented internally. Moreover, once redescription has taken place and explicit representations become manipulable, children can introduce violations to their data-driven, veridical descriptions of the world – violations which allow, for instance, for pretend play, false belief, and the use of counterfactuals. (Karmiloff-Smith, 1994, pp. 700-701)

Beyond E1 children's knowledge becomes more explicit and thus more flexible and manipulable. At level E2 children are supposed to have conscious, but not verbal, access to knowledge components. At level E3 knowledge elements become verbally explicit as well as enabling children to pass their knowledge components to each other. Karmiloff-Smith (1994, p. 701) admits that she generally treats levels E2 and E3 together because at that time she had not yet researched the distinction. A couple of concrete examples help to see what Karmiloff-Smith has in mind in laying out her representational redescription theory.

Clark and Karmiloff-Smith (1993) suggest that one example of representational redescription can be seen in children's understanding of the term *word*. Under their reading

---

[151] As Karmiloff-Smith forcefully points out, the pattern is repeated in all of these developing skills. However the pattern unfolds independently for each particular domain often at quite different ages. Thus, Karmiloff-Smith's domain specific approach is formulated in opposition to the domain-general framework of Piagetian theory. See Nelson (1996) for a moderate domain-general critique of Karmiloff-Smith.

[152] Karmiloff-Smith (1992, 1994) points out that the increasing explicitness of children's knowledge is not driven by the failure or lack of success of implicit strategies, but rather by achieving a *behavioural mastery* at the implicit level.

children initially have an implicit understanding of what a word is and this is evidenced by the fact that they can produce and comprehend words in a compositional manner. As they grow older this implicit understanding is representationally redescribed so that they can come to express this knowledge[153]. Karmiloff-Smith and her colleagues asked children between 3 and 7 years of age to engage in two different tasks. All of the children could produce multi-word sentences containing open- (roughly nouns, verbs, adjectives) and closed-classed (prepositions, articles, etc.) words. In the first task the experimenter read a story and occasionally paused mid sentence, whereupon the child was asked to repeat the last word that had been uttered. In the second task a number of open- and closed-class words were orally presented to the children and they were asked to say whether they were words or not. Three to 4.5 year olds could not do either task. Four and a half to 5 year olds could do the first task but not the second task. Usually children who failed the second task did not identify closed-class words as being words. Five to 7 year olds could do both tasks. Karmiloff-Smith takes these results to be supportive of the idea that at 4.5 to 5 years and 5 to 7 years the implicit knowledge (level I) about wordhood is representationally redescribed. That is, word knowledge is made more accessible and useable by the child for other kinds of activities besides the ones in which the distinctions were originally acquired. So the 4.5 to 5 year olds had redescribed their knowledge of what constitutes a word to an E1 level (that is they could extract information signalling whether a structure is a word within a partially on-line task) and that 5 to 7 year olds' cognitive systems had representationally redescribed their knowledge at a higher, more explicit, level (E2 or higher) where they can be directly questioned about the wordhood of different symbolic structures.

Karmiloff-Smith (1992) has shown similar increases in children's ability to explicitly decompose and reuse originally implicit knowledge in other domains. In one study she asked children to draw make-believe people and houses. She found that as children grew older they gained more explicit control over the elements in their drawings, being able to construct make-believe drawings from separate elements in normal drawings (e.g., putting aeroplane wings on a person). Younger children, on the other hand, produced rather stereotyped whole-drawings whose make-believe elements involved changes in size or

---

[153] Olson (1996) argues that people understand word boundaries explicitly only when they become literate. He presents several pieces of relevant research to support his claim. I am inclined to think that Olson's claim is more likely to be the case than Karmiloff-Smith's purely endogenous representational redescription approach. Nonetheless Karmiloff-Smith's experiment nicely illustrates what she means by representational redescription. It is also possible that one may want to argue a modified version of the representational redescription framework that explicitly includes exogenous factors such as an education within a literate culture.

shape rather than the juxtaposition of what are usually unrelated elements. In other words, younger children seemed to have great difficulty decomposing and reusing elements from their more holistic, interwoven, implicit procedural knowledge.

Karmiloff-Smith's interpretations of developing behaviours in terms of her representational-redescription model have been criticised by some for reading more explicit knowledge use into the data/phenomena than is warranted and of focusing too much on internal rather than external mechanisms (Bechtel, 1993b; Hendriks-Jansen, 1996). However, the general claim that Clark and Karmiloff-Smith make still stands. ANNs, as currently conceived, do not suffice as mechanisms that can enable advanced cognition. The ability to disembed knowledge from the structures that allow our ongoing cognitive activity is absolutely crucial to our ability to decouple from the immediate environment and think about things 'off-line' (cf. Grush, 1997). Yet there is nothing within the basic connectionist architecture itself that seems to endow cognizers with the ability to 'think off-line', because ANNs are implicit, environmentally embedded, devices rather than decoupled self-sufficient simulators.

Augmenting First-Order Connectionist Networks

Clark and Karmiloff-Smith (1993) suggest a number of ways that ANNs could be augmented in order to gain this sort of ability. The first is to give up on connectionism and embrace traditional symbolism. Of course, connectionists would be loathe to do this, given the acknowledged strengths of the approach. The interactionist is also unlikely to want to endorse the classical view, given its problems and anti-interactionist implications. The second is to embrace some sort of connectionist/symbolist compromise. Clark and Karmiloff-Smith discuss a number of existing hybrid architectures. However, this is no good for the ESD interactionist because it gives up the fight for an alternative explanation of cognition and also leaves unresolved the problems of cognitivism. The third alternative is to adopt some kind of 'second-order' connectionist solution. Clark and Karmiloff-Smith discuss Chalmers' (1990) use of the RAAM architecture (Pollack, 1990) which compresses activity patterns and passes it on to a second neural network that transforms the compressed data, according to some rule expressed in the dynamics of the second ANN, and then decompresses it as a new activity pattern (see also Bechtel, 1996). However, such a system still deals in emergent, implicit rules (in the transformation network) which cannot be transported or manipulated. Thus, no kind of second-order solution seems to currently exist.

Clark and Karmiloff-Smith clearly hope that some sort of second-order connectionist network can be created that fulfils all of their flexibility, transportability, and manipulability criteria but acknowledge that nothing of the sort currently exists. However, even if such a network did exist, what would be achieved is a thoroughly in-the-head

system for *implementing* just the kind of physical symbol system that Fodor, Pylyshyn, and other classicists argue must be the case.

The Appeal to External Symbol Systems

Clark and Karmiloff-Smith discuss another possible alternative that involves the use of *external* symbol structures (i.e., language and other symbol systems) as an environmental scaffold for enabling the advanced cognitive abilities that they seek to explain. However, they quickly dismiss this explanatory strategy on the grounds that children seem to exhibit explicit representational redescription abilities (level E1) without being able to linguistically express, manipulate, and use that knowledge (level E3). A number of commentators argue that such a dismissal may be premature (e.g., Bechtel, 1993b; Dennett, 1993; Hendriks-Jansen, 1996). Bechtel (1993b) argues that Karmiloff-Smith's examples do not unambiguously demonstrate that representational redescription is actually going on at the E1 level. He cleverly recasts each of the developmental changes noted by Karmiloff-Smith in terms of the "changing interactions between a cognitive system and external representations [e.g., utterances of words, drawings] rather than positing a process of internal representational redescription and operation upon that." (Bechtel, 1993b, p. 537). In each case the puzzle of how a network-like system could cope with the increasing 'explicitation' of knowledge is accounted for by off-loading the problem onto environmental structures which can be inspected by a first-order type system. Bechtel (1993b) goes on to say that

> I am concerned that in explaining the behavioral changes Clark and Karmiloff-Smith are constructing an excessively mentalistic story which will require a powerful cognitive homunculus that performs computations upon representations. I agree with them that more than first-order connectionist networks are required to explain human cognition, but I am concerned about the nature of what more is posited. ... I am suspicious of positing within the system the capacity to 'analyze' its own activity, for that seems to introduce an analyzer. Once a system has learned to generate and respond to external symbols such as those of natural language, then it does not seem problematic to talk of a system constructing an analysis of itself (perhaps only in private speech) and responding to it. (pp. 537-538)

I believe that Bechtel is on the right track here. Embodied and situated neural networks should remain as implicit, embedded-in-the-present, pattern mapping systems with the ability to pull off the kinds of modest time-space distancing mentioned earlier. After all, neural networks (that are implanted in a body, that provide the substrate for a subsumption-like architecture of behaviour layers, and that realise perceptual instruments, action generation processes, and neural control structures) seem to be able to, at least potentially, model nearly all of the basic cognitive activity exhibited by non-human animals as well as a large and significant amount of human cognition. This is no small accomplishment. Human advanced cognitive abilities obviously demand another implementational story. Instead of trying to push ANNs to some kind of second-order level that risks reincarnating

the problematic classical symbolic vision, it seems sensible to pay close attention to the structure of the complex social and symbolic environments in which advanced cognizers live. Surely this 'special input' may well play a significant role in supplementing basic cognitive abilities so that more radical kinds of time-space distancing are possible – distancing that involves forays into imagined and hypothetical environments as well as altering the way agents relate to the present, future, and the past. It is to these issues that we turn in the next chapter.

8. From Basic to Advanced Cognition

The doorway into this virtual world was opened to us alone by the evolution of language, because language is not merely a mode of communication, it is also the outward expression of an unusual mode of thought - symbolic representation. Without symbolization the entire virtual world ... is out of reach: inconceivable

Terence Deacon (1997, p. 22).

Introduction

The autonomous systems framework sketched in the last chapter accounts for, what I have called, *basic cognitive time-space distancing abilities*. Animals are not merely stimulus-response engines that are driven entirely by the 'projectable properties' in the local environment. Most animals are more closely tied to the nuances of their local environments than human beings. The objective of this chapter is to say something about what underpins this difference. Moreover, my goal is to argue that an interactionist approach can account for this difference without succumbing to the need to postulate a subpersonal representational explanation. Indeed, I believe that by comparison to the usual cognitivist story, interactionism's explanatory conservativism encourages a more sophisticated (and unfortunately much more complicated) understanding of how human symbolic cognition arises. The interactionist story is *necessarily* social, cultural, developmental, and evolutionary in nature. This radically expands the range of phenomena that need to be understood before a deeper understanding of human cognition can be had.

The Representation Hungry Problem

Here is the problem. As we have seen, it looks increasingly likely that within a sophisticated interactionist framework, we can show how an autonomous system, such as an animal or a mobile robot, can act intelligently with respect to things in its local environment. Workers in situated robotics have created subsumption architecture-based mobile robots that can make their way in real-world environments in a robust, timely, adaptive, useful, and intelligent-looking manner, all without the need to construct models of the environment or access explicitly laid down goals or plans (Clark & Toribio, 1994; Hendriks-Jansen, 1996). These situated mobots work in a distinctly un-cognitivist manner. They are sensitive to their surroundings (they 'perceive'), can demonstrate adaptive movements in a wide variety of circumstances (they can 'behave rationally'), and can come to anticipate events and streamline their actions (they can 'learn from experience'). In sum these critters are simple versions of insects like earwigs and other relatively context-bound creatures (Kirsh, 1991/1996).

However we know that many animals, especially many mammals and birds, are capable of coordinating their behaviour with elements of the surroundings with which they are not in

direct causal contact. I take it as uncontroversial that animals of all kinds are capable of incredible time-space distancing (TSD) abilities that deserve to be thought of as cognitive abilities. These abilities include the navigational and communicative skills of honey bees, the capacious memory abilities of birds like Clark's nutcracker, the complex alliance-tracking skills of monkeys, and the incredible migration capacity of animals like the arctic knot. Animals, it seems, are not merely reactive and situation-bound in a strict sense. In light of this fact Kirsh (1991/1996) has argued at length that the possession of non-representational, situated, subsumption-like architectures may be able to explain the behaviour of simple creatures like earwigs but not that of more sophisticated animals such as mammals and birds. Kirsh lists a number of phenomena that he believes require a conceptual/representational explanation. These include unifying perceptions into equivalence classes, making predictions, dealing with things out of perceptual contact, understanding things objectively, inference and predication, problem-solving, and engaging in creative endeavours. Similarly, as noted in Chapter 1, Clark and Toribio (1994) argue that without representation we cannot deal in cognitive abilities where the "problem involves reasoning about absent, non-existent, or counterfactual states of affairs" or the agent needs "to be selectively sensitive to parameters whose ambient physical manifestations are complex and unruly (for example, open endedly disjunctive)." (p. 419). More specifically Clark and Toribio argue that we require a representational explanation for dealing with:

1. Situations where there is an obvious absence of an environmental signal necessary for 'reasoning' about places, events, or individuals that are hidden, spatially distant, or that exist only in the past, or in the future. This is essence of what I have called *time-space distancing*.

2. Situations where behaviour seems to be coordinated with respect to non-existent or counter-factual states of affairs. Phenomena such as imagination, misrepresentation, and deception fall into this category.

3. Situations where animals seem to *compress* or *dilate* perceivable stimuli in ways which do not depend upon perceptual similarity and difference. This is what is usually meant by the term *conceptual equivalence*.

One could argue that there is no evidence that any non-human animals possess any of these capacities in the sense that humans do. For instance, on reviewing the evidence Donald (1991, 1998, 1999) argues that even apes – animals that are surely amongst the most cognitively sophisticated in the world – inhabit a purely *episodic culture* which is characterised as primarily unreflective, concrete, and situation-bound in nature (see also Heyes, 1998; Lock & Colombo, 1996; Nelson, 1996). According to Donald modern apes possess a complex event perception ability and an episodic memory that is reliant on

situational triggers. Such abilities are evident when, for instance, chimpanzees 'recognise' specific individuals, places, events, and actions and can modulate their behaviour accordingly. In this sense episodic ability is a powerful kind of procedural skill (Nelson, 1996) and is thus, at least conceivably, achievable by some kind of advanced, embodied, first-order connectionist network. In a discussion of Norman and Shallice's (1986) theory of the *supervisory attention system* Dennett (1998) also points out that non-humans rarely exercise the advanced cognitive skills that are supposed to flow from high-powered central cognitive mechanisms – skills such as planning, troubleshooting, dealing with novelty, dealing with danger, and overcoming habits. Dennett (1998) argues that these skills are not reliably available to animals without language even though "[t]here are plenty of familiar anecdotes proclaiming that birds and mammals – at least – exhibit these talents on occasion." (p. 294). He makes the point that we should be impressed by the fact that examples of these skills in animals "recount remarkable and impressive cases" and that "[i]n addition to the anecdotes of glory, there is evidence, both experimental and anecdotal, of their widespread failure to rise to challenges of this sort." (p. 294).

Despite these observations many animals *do* seem to be able to adaptively modulate their behaviour with regard to aspects of the world with which they do not have causal contact. There is an important sense in which many animals strive for 'goals' which involve distant places or objects and events (not currently available). These cases of *anticipatory behaviour* (Keijzer, 1997, 1998a) seem to require the postulation of in-the-head representations of these non-local entities (models) and plans for obtaining them (what Keijzer [1997, 1998a] calls *representations that specify what to do*). The challenge for a non-representational interactionist framework is to account for these kinds of basic time-space distancing without requiring a subpersonal representational explanation.

Explaining Basic Cognitive Capacities in Non-humans

I think that there are at least two ways in which basic long-distance, time-space distancing skills can be accounted for in a non-representational interactionist framework. The first way is, as was suggested in the previous chapter, to exploit the temporal pattern matching powers of recurrent networks (Elman, 1990). If animals incorporate something like the kind of neural organisation seen in recurrent networks, that is, a set of context nodes that feed activation values from a previous time-step into the current network dynamics, and, if it makes sense to think of neural networks as non-representational systems, as I think it does, then it is quite possible that we can think of such abilities within a sophisticated basic cognitive interactionist framework.

Secondly, it may well be possible that non-recurrent systems might exhibit temporally-sensitive behaviour by simply learning about, or evolving mechanisms to deal with, the causal contingencies that exist between different parts of relevant environmental events.

For instance, as noted in the last chapter, Pfeifer and Verschure's (1992a, 1992b) robot Lola, manages to do things that the builders imagined would require memory (i.e., taking account of things that are currently not the case), with simple feedforward neural networks. In a sense Lola 'sees' the past in the present.

Keijzer (1997, 1998a) has attempted to make explicit the substance of these ideas in arguing that anticipatory behaviour can be explained within an extended interactionist framework where large-scale dynamics can be modulated via the control of small-scale local dynamics. Keijzer's idea is a complex one and requires an appreciation of dynamical systems theoretic concepts. As noted in chapter 4, he argues that organisms capable of anticipatory behaviour should be thought of as possessing neurally-realised *internal control parameters* (ICPs) that adaptively perturb behavioural patterns at higher levels of organisation. The idea is that the ICP need not contain a detailed context-independent content that specifies what an animal should do (i.e., an ICP is not a plan-like representation). Rather an ICP is simply a 'trigger' for initiating a reorganisation of behavioural dynamics at multiple scales of organisation.

> The neural system modulates short term organism-environment interactions. These short-term interactions in turn modulate happenings in the environment on a longer time-scale, and so on. Large-scale order results from intrinsic self-organizing tendencies at that scale. This order is manipulated by changing control parameters within the microscopy. It is not specified beforehand within the microscopy. ... The microscopy 'uses' the intrinsic order at larger scales to generate the trajectory. (Keijzer, 1997, p. 231).

The mosquito's blood-finding behaviour provides a simple example. The internal mechanisms (ICPs) of the mosquito act only with respect to the local environment (the presence of CO_2) but the 'natural dynamics' of CO_2 is related to other events on a larger-scale (i.e., respiration events in animals). By modulating behaviour with respect to local events, the mosquito also modulates its behaviour with respect to larger-scale (temporally and spatially larger) events. However, there is nothing in the internal mechanism that needs to possess a contentful structure that relates to anything about the larger-scale events. That is dealt with entirely by the lawful relations that hold between differently scaled environmental phenomena. The 'content', if it is anywhere, is distributed across the agent-environment system[154]. So Keijzer suggests that animals have evolved regulatory ICP networks (neural/bodily subsystems) that enable behavioural modulation over large scales by creating local changes which have adaptive long distance consequences:

---

[154] And, as has been noted in earlier chapters, the notion of *content* does not really fit such a view. Content implies a content holder separate from the thing that the content is about. In the distributed interactionist view there is no 'information in the head' just an adaptive capacity for sensitive coordination.

> [A]nticipatory behavior occurs when a behavioral system is capable of generating regular trajectories, which lead to a sufficient macroscopic order. Given a capacity for self-organized behavioral regularities at a short time-scale, a regulatory network which modulates a set of control parameters will be able to guide this process over longer time-scales in a way that makes it relatively independent of the immediate environment. It will enable the system to achieve long-term goals. An extended interactionist account of adaptive behavior can thus deal with problems which formerly necessitated representational specifications. (Keijzer, 1997, p. 234)

In an important sense Keijzer here challenges the idea that the term *situated* is synonymous with the term *stimulus-bound.* The latter term implies that behaviour can only be coordinated with respect to something with which the agent is currently in direct causal contact. Situatedness, on the other hand, refers to being sensitive to (i.e., being able to modulate behaviour with respect to) some aspect of the environment and this may be possible via the causal consequences of the distant environmental state on more proximal parts of the world. So any behaviour that can be cashed out in terms of modulation with respect to a sequence of environmental 'clues' and relations (a trackable gradient as suggested by Christensen & Hooker [2000, in press]) constitutes a basic cognitive activity.

Language, Symbolic Skill, and Humans' Radical Time-Space Distancing Ability

Despite the fact that Keijzer's ICP hypothesis seems a promising way to extend the interactionist approach to account for anticipatory behaviours such as long-distance migration, temporally extended food searches, and abilities to relocate particular places in the environment, it does not tell us the whole story about the powerful time-space distancing abilities of human beings. Keijzer happily admits that he has deliberately restricted his focus to *behaviour* (roughly what I have been calling basic cognitive activity) rather than high-level cognition. So the ICP hypothesis does not directly tackle the kinds of phenomena that cognitive psychologists commonly address – things like reasoning, problem-solving, and explicit memory. And it is just these kinds of *advanced cognitive abilities* that seem to cry out for representational explanation. We can get a better grip on the contrast between basic and advanced cognitive abilities by briefly examining cognitive developmental phenomena.

Human Cognitive Development

Many, perhaps most, theories of cognitive development can be understood as portraying the child moving through a series of three basic phases of cognitive functioning. The child begins as a situated basic cognizer, albeit one embedded in a complex sociocultural environment, and then moves through a transitional stage where nascent symbolic abilities develop until, about the age of 4 years, they begin to behave in a recognisably advanced manner.

Basic Cognitive Period

During the first period of development the infant shows a developing ability to make sense of their immediate environment. They develop the kinds of abilities that all animals possess for survival in the world. These include the ability to perceive the affordances of the surroundings and to act adaptively with regard to what is perceived. Piaget (1970) famously characterised this period as a *sensorimotor* stage of development. He believed that the newborn needs to learn the most elementary of skills such as distinguishing self from non-self, and learning about the basic physical characteristics of objects, such as persistence and substance, by actively exploring themselves and their surroundings. However these ideas have not stood up to empirical scrutiny. It has become clear that even very young infants seem to be quite well equipped for making sense of the world as a three-dimensional place and themselves as actors in it. Other abilities develop quickly such as deferred imitation (Meltzoff, 1988; Meltzoff & Moore, 1977), tracking objects and events through brief displacements (Baillargeon, 1986, 1993; Diamond, 1985, 1988), and distinguishing phenomena such as human speech sounds (see Bates, Thal, & Janowsky, 1992). Thus many modern developmental theorists have objected to the characterisation of the infant as a sensorimotor creature (e.g., Mandler, 1992; Meltzoff & Moore, 1998; Spelke, 1998) and claim that infants are actually quite complex innately *conceptual* agents. Infants rapidly develop the capacity to bridge-time (what I have called *diachronic pattern mapping*), to 'remember' from only one or a few experiences, and to recognise an object using different modalities (crossmodal integration). These findings have, for instance, led Suddendorf (1999) to the cognitivist conclusion that, up until 18 months of age, children must be in possession of a 'single updating *model* of the world' because they do not merely feed off the 'projectable properties' available in the local environment (see also Perner, 1991). Instead they *seem to be* collecting stimuli into equivalence classes ('concepts') and using these in an off-line manner when they track objects and imitate the actions that they have seen others perform (Meltzoff & Moore, 1998). However others have argued that many of the claims of these, often strongly nativist, representationalist claims about young infants' cognitive abilities are not supported by current experimental findings, and that infants should be considered to be essentially situated, sensorimotor beings (Fisher & Biddell, 1991; Haith, 1998). Rutkowska (1994) has even explicitly argued that infants should be thought of as complex kinds of Brooksian mobots – organic machines whose behaviour derives from the interaction of three kinds of neural system (perceptual processes, behavioural/motor processes, and action programs) that softly assemble independent 'reflexes', and the embedding environment (see also Hendriks-Jansen, 1996). She believes that infant abilities, including those that exhibit anticipation and expectation, can be modelled without needing explicit goals or hierarchical control. For instance, she claims that anticipation of an occluded object's trajectory "develops from the

concatenation of two initially independent processes: novel attention to the object's kinetic occlusion; and turning (necessarily in the direction of its unseen movement) to (re)fixate the object." (Rutkowska, 1994, p. 184).

So infants, among other basic cognizers, perceive patterns that stretch over (usually short) spans of space and periods of time. They can enter into a perceptual relationship with an event given only small 'cues'. That is, they can 'enter' a pattern at many points and can use their exploratory abilities to provide fuller information for more completely specifying the nature of the event. They can thus modulate their activity in terms of what they perceive. There is no need to postulate internal inferential mechanisms that make use of mental representations. What often looks to an observer to be complex, canny, and adaptive behaviour that results from a detailed understanding of the layout of the environment (a mental model) operated on by a planning system, actually derives from the interaction of simple behavioural layers with the structure of the local environment. My claim here is that *all* of the behaviour of nonhumans and of human infants can be explained in terms of such non-representational basic cognitive mechanisms, processes, and skills (with the possible exception of enculturated apes).

The Transitional Period

The second phase of cognitive development is a transitory phase between the first context-bound phase and the third symbolic and representational phase. Around the age of two years children seem to acquire, what Piaget (1951, 1970) referred to as *the symbolic function* (see also Russell, 1996). About this time children begin to develop a number of simple abilities which involve an ability to "decouple action from its immediate pragmatic function ..." (Nelson, 1996, p. 102). Suddendorf (1999) suggests that this ability for "mental detachment from the immediate present" (p. 227) derives from an ability to entertain additional models ('interpretations') of the environment to the 'single updating model' that they possess from birth (see also Leslie, 1987). This reveals itself in a wide variety of new abilities that seem to involve an ability to juxtapose an alternative 'decoupled' view of a situation (e.g., a desired future state, an expected state, a remembered state, etc.) with one's immediate perception of the situation. Thus children of this age exhibit 'insight' and planning (juxtaposing a desired goal state with the current state of the environment), Piaget's stage 6 object permanence (the capacity to remember that an object exists after it is hidden), pretence (the ability to simultaneously 'hold in mind' the actual nature of an object, such as a toy, and a representation of what one is pretending it to be), the awareness of another's perspective of a situation, and complex imitation (the ability for the precise practicing and rehearsal of actions [see also Donald,

1998, 1999])[155]. Although the 'holding in mind two models of the world' account of these abilities is popular (e.g., Leslie, 1987; Olson, 1989, 1993; Perner, 1991; Suddendorf, 1991), it is by no means the only way of understanding these developments. Nelson (1996) argues that the ability to decouple does not so much arise from the development of a more complex 'neurocognitive architecture', but from learning about the way certain events unfold in the child's social world. She speculates, for instance, that parents or older siblings jointly construct pretence with children by engaging in activities where familiar routines or events can be played out in an out-of-context manner. For example, a parent might pretend that stuffed toy animals are going to bite, kiss, or hug the child. From experiences like this the child gains the insight that some activities are pretend ones. In a sense Nelson argues that children come to understand how, and in what ways, they can generalise particular 'real world' activities like bathing, sleeping, and so on to other pretend actions (e.g., lying down and pretending to sleep) and objects (e.g., using a doll as a prop for a familiar activity). In other words, children learn to 'stretch' the conditions under which they will deploy their basic situated activity skills.

Despite possessing these early off-line skills, children of these ages are still unable to engage in adult-like off-line cognizing. Children of between about one and a half and four years largely rely upon caregivers to scaffold their skills (Nelson, 1996). Without this help these children remain unable to exercise much of an ability to transcend the here-and-now. Children younger than four do not have adult-like explicit memories. Memories of past episodes do not easily come to mind and the memorial narratives that they produce are often fragmentary in nature. They are also often confused about the source of their memories. In fact children of this age do not possess a long-lasting narrative-based autobiographical memory at all (Fivush & Schwarzmueller, 1998; Nelson, 1993, 1996)[156]. In other words, they do not exhibit the ability to voluntarily and independently 'retrieve' even relatively recent life events and episodes. They cannot control their memories in an

---

[155] Russell (1996) feels that it is important to distinguish the precise and complex ability to imitate that emerges at about age two from the more basic deferred imitation exhibited by infants (Meltzoff, 1988; Meltzoff & Moore, 1977). He suggests that early deferred imitation might be "described more parsimoniously as 'remembering what actions a novel object will afford' ..." and that the more complex imitation skills that Piaget (1951) had in mind "were often of the child's imitation of how another person behaved (rather than what was done with an object) – such as the imitation of a playmate throwing a temper tantrum the previous day. If we mean *this* by 'deferred imitation' then there are no strong impediments to saying it arrives some time in the middle of the second year." (p. 160).

[156] Interestingly the time of onset of autobiographical memory corresponds to the ending of childhood amnesia – the inability for adults to remember events that occurred before 3.5 to 4 years of age (Fivush & Schwarzmueller, 1998; Howe, 1998; Howe & Courage, 1993, 1997; Meltzoff, 1995; Nelson, 1993, 1996).

adult-like fashion[157]. Similarly these children's categorisation skills are mostly perceptually-based rather than hierarchically and logically-based. They do not seem to possess the strong compression and dilation skills that Clark and Toribio (1994) argue require a representational explanation. 'Transitory children' do not fully appreciate adult cultural temporal concepts (such as *before*, *first*, and *yesterday*). Although many children of this age can count, they do not understand what counting is for. Despite having just counted all of their toys the pre-four-year-old will not then know how to answer the question "How many toys do you have?" (Deheane, 1998; Wynn, 1992). And, perhaps most famously, pre-four-year-olds regularly fail *false belief tasks*, where the child needs to appreciate that another's beliefs (or even their own past beliefs) may differ from their own (Wimmer & Perner, 1983), and *appearance-reality tasks*, where the child does not seem to be able to acknowledge that an object can look like one thing but actually be another (e.g., a sponge painted to look like a rock) (Flavell, Green, & Flavell, 1986). In all of these cases two to four year old children show that they cannot quite fully transcend their basic cognitive situation-embeddedness. They do not possess the independent ability to easily and voluntarily think off-line, when it comes to dealing with distant and imaginary environments, although they show signs of applying their basic cognitive skills to situations in new and unusual ways.

The Early Advanced Cognition Period

The third phase of development signals the *beginning* of advanced cognitive activity. Somewhere between the ages of three and a half and five years children develop the ability to think about the past, the future, the out-of-context, and the imaginary in a reasonably adult-like form. This is evidenced in a variety of domains. Children of this age begin to develop a long lasting narrative-based autobiographical memory, they can construct fluent personal episodic accounts and create short made-up stories. They begin to show evidence of using logically-based categories instead of perceptually based ones. They come to appreciate cultural and technological notions of time and temporality. They develop a sense of self as different from others and a new understanding of the intentional states of others emerges. They are able to formulate plans and long term goals for future action, recollect specific events in a coherent and explicit manner, make accurate predictions of people's behaviour based on what is known about their beliefs, engage in lying and

---

[157] The Flavells and their co-workers have shown that those I have been calling 'transitory children' generally show little in the way of *metacognitive skill*. That is, they do not show an awareness of, much less a control of, the what adults refer to as the flow of consciousness, inner speech, or imagery. In addition, they exhibit little in the way of metamemory skill – the use of deliberate strategies for organising and controlling

deception, effectively restrain themselves from immediate impulsive action, make sense of complex narratives, amongst a variety of skills that require an ability to set a virtual situation apart from a real one (Nelson, 1996; Suddendorf, 1999). To be sure these skills reach higher levels of sophistication until the age of 10 and beyond, but children of about four are recognisably advanced in their cognizing[158]. Lying behind all of these skills seems to be the basic ability to bracket beliefs and points-of-view from what is or what seems to be the case. Children at this stage of development appreciate the power of language as a 'virtual environment builder' and use this understanding when they comprehend and produce utterances. Suddendorf (1999) argues that these abilities derive from an ability to *metarepresent*; to understand representations *as* representations and to appreciate how representations relate to reality:

> With metarepresentation, then, representational relations can be tagged with predicates (e.g. your view; my memory), and the individual can now simultaneously entertain several distinct representations of the same object or event in reality without running into paradoxical conflicts. ... With metamind the child can appreciate representations as representations and can entertain various conflicting representations of the same object or event. ... Metamind enables the individual to entertain various ways of looking at the same thing – representing what it is, looks like, was, could be, should be, and so on. ... At times, metamind can 'wander off', as it were, and entertain a variety of propositions. It is the stage for complex reasoning, considering 'what if?', theorizing, reconstructing the past, and planning the future. *Metamind enables the cognitive apparatus to function off-line.* (Suddendorf, 1999, p. 235, emphasis added)

Ultimately advanced cognition comes down to the ability to create 'virtual environments' that enable a peculiar kind of flexibility – an ability to 'make oneself do things out of context' rather than merely be sensitive to the environment (local or distant). One must be careful here, however, as it may be the case that behaviours generated by these two different 'cognitive systems' may not be empirically distinguishable from each other. When simply observing behaviour in an isolated context it is difficult to tell whether the activity is being coordinated through the tracking of environmental gradients or via an off-line use of virtual environments.

Advanced Cognition and Imagery

---

memory such as rehearsal of lists (Flavell, Green, and Flavell, 1993, 1995, 1996; Flavell, Green, Flavell, & Grossman, 1997).

[158] Developmental theories often include a later stage of development where formal, theoretic, and logical cognition is understood to emerge somewhere around the age of 11 or 12. Piaget (1970) called this phase of development the *formal operational stage*. For a detailed neo-Piagetian theory of the cognitive developments in middle childhood see Case (1985). K. Gibson (1996) also provides a summary of neo-Piagetian thinking. Both Scheerer (1996) and Olson (1994, 1996) argue that complex human cognition may require literacy.

The Recruitment Hypothesis

I propose that we understand advanced cognitive abilities as grounded in a capacity to use our basic, situated cognitive abilities in an 'off-line manner'. This is achieved by being able to, in a sense, create a kind of virtual environment into which we can plug our basic cognitive mechanisms. Moreover, I want to claim that this recruitment ability rests, not on a built-in neural emulation machine, but on the way complex social and linguistic environments affect our basic cognitive architecture. This does not mean that our brains are not well-designed to facilitate the emergence of advanced cognizing. Rather, I want to suggest, in true interactionist fashion, that the structure that enforces and enables advanced cognitive abilities is situated in the human environment, especially in its social and linguistic structures. I will develop my, admittedly speculative, ideas in the following way:

Humans and other animals are complex basic cognition machines designed to negotiate and use local environments often in the service of long distance goals (anticipatory behaviour). As I have argued in the previous chapters, the fundamental unit of basic cognition is the perception-action cycle and the agent-side component of the cycle is largely realised by a regulative, coordinative, and integrative nervous system. I call this the *basic cognition package*.

Advanced cognition involves the ability to recruit the mechanisms and reuse the abilities of the basic cognition package. We see this phenomenon most clearly in mental (visual, auditory, and motor) imagery. We already know from imagery research that human beings use their basic perception-action abilities, and recruit the neural mechanisms that underpin those abilities, in the production of off-line imaging. Imaging can be thought of as an ability to entertain realities and possibilities beyond the immediate here and now, or as the ability to generate a virtual environment with respect to which we can modulate our actions. So it possible to think of advanced cognition as imagery use in a broad sense.

People are not born with the ability to image in a controlled fashion. Instead we come to learn to control and reuse our basic cognitive package through an immersion and education within a human social and communicative environment. At the root of learning to reuse imagery is our ability to use natural language and other public symbol systems.

The next section, *Imagery as the Reuse of Basic Perception-Action Systems*, examines whether imagery can be understood as a kind of decoupled version of perception and action. The subsequent section, *Learning to Control Imagery*, speculates about the ways in which immersion in a social and linguistic environment may enable this decoupling.

Imagery as the Reuse of Basic Perception-Action Systems

An interactionist theory of cognition is committed to a position that eschews the need for representations that serve as inner models of the world or inner plans for action. This

means that if we want to develop an interactionist explanation of advanced cognition we need to avoid postulating the existence of neurocognitive modules or circuits that *specifically* enable advanced cognitive abilities. That is, we need to explain advanced cognition in terms of some redeployment or reuse of the skills and mechanisms that underpin basic cognition. This does not mean that the brains of advanced cognizers will be identical in structure and function to those of basic cognizers, only that advanced cognition does not require an evolved, neurally-realised, inner modelling device. This, I believe, is more or less what Vygotskians have in mind in claiming that "[i]f one decomposes a higher mental function into its constituent parts, one finds nothing but the natural, lower skills." (Kozulin, 1986, p. xxv). Fortunately there is good evidence to suggest that at least *some* off-line cognitive activities make use of basic, on-line neural resources.

Evidence for Recruitment in Imagery

Many of the observations made in imagery research support the view that the neural mechanisms underlying perception and action are used when people imagine particular objects, events, and actions. Mental imagery research has primarily focused upon its perceptual aspect. For many of us experiences of visual and auditory images have a very similar feeling to our perceptual experiences. They are different enough however for most of us to know that they are not identical. Experimental research also provides a good deal of support for the idea that imagery involves activity in perceptual mechanisms. For instance, it has been widely shown that imagery of one modality can interfere with perceptual tasks in that same modality (and vice versa) (e.g., Craver-Lemley & Reeves, 1987; Craver-Lemley, Arterberry, & Reeves, 1997; Smith, Reisberg, & Wilson, 1992). These findings have been taken to imply that the many of the same neural and bodily resources are required for carrying out both kinds of activity. Neuroscientific research has further supported this idea by showing that, in normal, non-brain damaged individuals, many of the regions in the brain that are active during perceptual activities are also active in imagery activities that parallel the perceptual tasks (on visual imagery see Farah, 1989; on motor imagery see Jeannerod, 1995; on auditory imagery see Mcguire, Silbersweig, Murray, David, Frackowiak, & Frith, 1996; Zatorre, Halpern, Perry, Meyer, & Evans, 1996). There is even evidence to suggest that people who suffer deficits in perception due to neural damage suffer parallel deficits in imagery. These parallels include colour-blindness, what-where dissociations, face agnosia, and visual neglect (Farah, 1989).

However, the evidence is not entirely unambiguous. Goldenberg, Müllbacher, and Nowak (1995) describe a patient who is completely cortically blind (near complete damage to the primary visual cortex) but was capable of providing appropriate answers to questions that are usually thought to require visual imaging. Similarly Thomas (1999) reports that Roland and Gulyás (1994) and Mellet et al. (1996) claim to have shown that no activity occurs in

the retinotopically mapped regions of the visual cortex when visually imaging and that such activity is more consistently associated with neural activity in non-retinotopically organised areas (see also Deposito, Detre, Aguirre, Stallcup, Alsop, Tippet, & Farah, 1997; these findings have been contested by Kosslyn, Thompson, Kim, & Alpert, 1995)[159]. This of course may not mean that there is *no* recruitment of perception structures in visual imagery, but this research should make us sceptical of claims that the entire neural circuitry of vision is utilised. Indeed most researchers believe that recruitment does occur, perhaps primarily in non-primary cortices such as the visual association cortex (Deposito et al., 1997).

Although mental imagery, especially visual imagery, is often taken to be a sort of out-of-context, off-line partner of *perception,* there is also plenty of evidence that imaging involves *motor* components (Annett, 1995; Kosslyn, Behrmann, & Jeannerod, 1995; Jeannerod, 1995). This is not only the case when one is imagining carrying out a physical activity such as reaching for an object (Georgopoulos, Lurito, Petrides, Schwartz, & Massey, 1989) or tying ones laces (Annett, 1995), but also when one is engaged in visual image transformations such as rotating an imagined object (Deutsch, Bourbon, Papanicolaou, & Eisenberg, 1988; Wexler, Kosslyn, & Berthoz, 1998). Annett (1995) presents research that shows that muscular activity occurs when people imagine how to perform a motor task and Jeannerod (1995) reviews a number of studies that show that imagining a limb moving increases the preparedness of the limb's reflex capacity in the same way that real movement does (but to a slightly lesser extent). Similar phenomena have been observed in tasks involving inner speech, which is widely taken to be a form of auditory imagery (Sokolov, 1968/1972; MacKay, 1992; Reisberg, 1992). Although evidence is against muscular activity being *necessary* for imagery[160] it is a common accompaniment in non-brain damaged individuals. Glenberg (1997) describes how bodily activity is associated with a wide variety of imaging activities:

> [I]magining a fearful situation evokes sweating, imagining a positive situation results in measurable activity in muscles associated with smiling, and imagining negative situations results in muscles associated with furrowing of the brow. There are analogous effects for imagery related to other perceptual/action systems.

---

[159] Part of the problem here must be due to the widespread use of the subtraction technique in PET, fMRI, and ERP research. The 'activations' that cognitive neuroscientists observe are made relative to one or more comparison conditions. This means that any observed activation is dependent upon the degree to which the tasks in the comparison conditions involve similar neural resources.

[160] The classic demonstration of this is Smith, Brown, Tolman, and Goodman's (1947) study where the unfortunate Smith had his entire musculature paralysed using D-tubocurarine and was kept breathing using artificial means. While paralysed Smith could observe, comprehend speech, retrieve memories, and think normally (see also Damasio, 1999, pp. 292-294).

Thus, in imaging a pendular motion, discharges in the eye muscles follow the appropriate frequency, in imaging bicep curls there are discharges in the biceps, and in imagining the taste of a favorite food there is an increase in saliva flow. (p. 5)

These findings suggest that imagery involves a kind of retro-activation of mechanisms that underpin motor activity as well as perception. In other words, imagery probably involves the 'priming' of the entire body to react as if it was really perceiving the imagined event or object.

In sum, then there exists considerable evidence in support of the view that basic cognitive structures used for making do in the local environment are redeployed when we entertain things 'off-line'. It is not a big leap to suggest that it may be our imagery abilities such as visualisation, mental practice, and inner speech that enable us to accomplish advanced cognitive tasks. Indeed much research shows that imagery can play an important role in laying down memories, recall, concrete and abstract reasoning, language comprehension, and the learning of new skills (Kosslyn et al., 1995). Within the framework I am developing here one can think of imagery primarily in terms of the ability to control and evoke basic cognitive abilities in out-of-context situations. In fact, perhaps ironically, this view of imagery plays down the experiential and conscious aspect of imagery and focuses upon the idea that imagery involves an ability to 'context hop' (Lock, 1999) or 'suppress' (Glenberg, 1997) – to get ourselves to do things that are not responses to natural immediate environmental goings on by putting ourselves in a 'virtual context'. Psychology and cognitive neuroscience have provided us with some fascinating facts about the relationships between imagery, perception and action, but it is important to think about how we might explain how basic perception-action structures may be recruited in the service of advanced cognitive abilities.

How Might Imagery Account for Advanced Cognition?

Cognitivist theorists generally explain imagery as the product of plugging a descriptive mental representation into the 'back end' of our normal perception-action apparatus to produce a surrogate for normal perceptual information (the image) (Kosslyn, 1994). This, as was briefly noted in chapter 2, is the approach that Grush (1997; see also Clark & Grush, 1997[161]) takes with regard to *mental representation*. He suggests that basic

[161] Clark (1997) provides a friendly critique of this notion that denies the need for the 'brain state' (that constitutes the representation) to be used 'off-line'. Clark's tinkering effectively reduces the power of this theory of representation by requiring it to make use of the vague idea that a representation is a representation if it has been designed (e.g., by natural selection) to transmit a particular kind of information to other parts of the 'cognitive system' (see also Chemero, 1998a, 1998b for a critique of the notion of representations as 'stand ins').

perceptual engagement with the local environment does *not* require a representational mechanism (but see Clark & Grush, 1997). Instead he claims that the neural activity associated with the energetic stimulation of the sensory surfaces by something in the environment constitutes a *presentation*. Presentations do not fulfil all of the criteria generally thought to exist for representations and are something like Dretske's (1988) indicators or type II representations. The issue of representation only arises, argues Grush, when something in the brain has to do the job of something in the world when in fact that thing is not currently in the environment and being sensed by the agent. That is, a representation is part of a system that *emulates* the perceptual input of an absent environmental object. A representation must be an internal content-preserving structure that is *decouple-able* from the real world and that can be plugged into the behaviour producing apparatus in the nervous system (see also Clark, 1997; Haugeland, 1991)[162]. In support of this claim Grush (1997) argues, and supplies some supporting evidence, that some activities (e.g., rapidly reaching for an object) require a motor emulator because neural feedback of proprioceptive information about the position of the arm is too slow to effectively modulate the trajectory of the moving arm. The emulator architecture can overcome this problem by presenting the bodily action systems with a representation of 'the way the world will be in a few moments', thereby allowing the agent to act earlier than would be possible if they were relying upon perceptual feedback to direct action.

For various reasons I believe that Grush's reaching examples do not provide convincing evidence for low-level emulation. Firstly, Grush (1997) admits that existing empirical evidence regarding motor activity, such as reaching, may be consistent with a feedback model. Secondly, even if it turns out that a feedback model is insufficient to account for rapid reaching, it is possible, I believe, for such a system to be implemented within the basic sort of pattern-mapping system I outlined above, for, as Chemero (1998a) notes "[d]espite the fact that the emulator is not hooked-up to incoming proprioceptive signals for a few milliseconds, the arm and the action it is undertaking is in no way *absent*." (p. 34). It seems plausible that a pattern-mapping system might be able to implement this kind of moderate anticipative activity. Finally, the motor emulator notion does not actually fulfil all of the criteria Grush (1997) requires of an emulator theory, particularly the need to be

---

[162] Within psychology a similar claim has been made by Leslie (1987; see also Olson, 1989, 1993; Russell, 1996; Suddendorf, 1999) to account for phenomena in developmental psychology relating to theory of mind research. Specifically Leslie hypothesises that children can only *pretend* when they can juxtapose a representation of an actual object (e.g., a banana) with, what Leslie calls, a *meta*representation of the thing that the object is taken to be (e.g., a telephone). To manage such a feat children must possess a decoupler – a device for removing a representation from its normal causal input-output functions and inserting it into an appropriate place in the information-processing sequence.

able to identify subcomponents within the representation bearer. Grush thus provides no compelling evidence for the idea that there exists an emulator as a subpersonal imagery mechanism in real animals. But if such a system does not exist, how do we explain how mental imagery works?

The Perceptual Activity Theory of Imagery

Thomas (1999) develops a theory of imagery that avoids a commitment to, what he calls, the *computational mentalism* of the quasi-pictorial theory of imagery (e.g., Kosslyn, 1994) and the description or propositional theory of imagery (e.g., Pylyshyn, 1991). He calls his approach a perceptual activity theory and which draws upon insights in active vision research (see chapter 3). In this view "perceptual learning is not viewed as a matter of storing descriptions (or pictures) of perceived scenes or objects, but as the continual updating and refining of procedures (or "schemata," see Neisser, 1976) that specify how to direct our attention most effectively in particular situations: how to efficiently examine and explore, and thus interpret, a scene or object of a certain type." (Thomas, 1999, p. 218). Imagery occurs when the perceptual instruments are retro-activated via the activation of neural structures controlling the coordination of the instruments (schemata):

> In this theory imagery is experienced when a schema that is not directly relevant to the exploration of the current environment is allowed at least partial control of the exploratory apparatus. We imagine, say, a cat by going through (some of) the motions of examining something and finding that it is a cat, even though there is no cat (and perhaps nothing relevant at all) there to be examined. *Imagining a cat is seeing nothing-in-particular as a cat.* (Thomas, 1999, p. 218, emphasis added)

When we view a cat in the environment our perceptual instruments are deployed to measure and test elements of relevant information about the cat. When we visually imagine a cat we activate those instruments to work in the same manner in the absence of any relevant environmental stimulation. Because we have previous experience of deploying our perceptual instruments in the perception of cats (and cat pictures, sculptures, cartoons, etc.) our brains have developed the dynamic neural connections that reflect the synchronic and diachronic activity of our perceptual instruments for many kinds of events, objects, and properties. Perhaps something like Edelman's (1989, 1992) re-entrant signalling is at work here in the generation of association patterns between instruments. Notice how this picture differs from Kosslyn's approach (1994) where modality-free *representations* in long term memory are translated into a perceptual medium in imagery. First of all, the 'schemata' in our neural control structures are 'instructions' (connections and weightings) for formatting instruments in particular conditions, not representations of things in the world. They are neural circuits that set off a self-organising pattern of neuromuscular activity that organises or primes the body into a activity pattern similar to that which occurs when perceiving an

environmental object or event[163]. Secondly, the perceptual activity view holds that "no end-product of perception, no inner picture or description is ever created. No *thing* in the brain is the percept or image." (Thomas, 1999, p. 218). Rather, imaging involves 'priming' the agent's body to act *as if* it was perceiving something[164]. The problem for Thomas' (1999) approach is in specifying how it is that "a schema that is not directly relevant to the exploration of the current environment is *allowed* at least partial control of the exploratory apparatus." (p. 218, emphasis added). Why, in other words, do we 'see' (or hear or feel) an object and event that is not the one in front of us? How is it that we can initiate activity that is not our 'natural' (situated, basic) response to a situation?

Learning to Control Imagery

The problem we face is the identification of a mechanism for enabling people to use their perception-action resources in an off-line manner. In other words, we are trying to explain how 'imaging' might be possible within an interactionist framework that eschews the need to postulate an internal representational trigger for the recruitment of perception-action structures. What mechanism can enable the off-line assembly of our perception-action resources? I want to argue that a command of language and other symbol systems can facilitate this possibility. But how can we develop this idea? After all, the dominant (Chomskian, generative grammatical) view of language within cognitive science merely treats language as an adjunct to thought – a kind of external expression and comprehension system bolted on to an independent central conceptual system (e.g., Chomsky, 1995; Pinker, 1994). Carruthers and Boucher (1998) refer to this as the *communicative* view of the relation between language and thought, a view in which language does not affect thought at all. Indeed in an important sense thought 'comes before' language. So, in this view, our advanced cognitive ability must be a feature of 'pure thought'. The answer, I believe, rests on combining what Bechtel (1993a, 1996) has called a *distributed view of language* or an *external symbol approach* with a roughly Vygotskian view of the regulatory powers of language (e.g., Berk, 1994; Diaz & Berk, 1992). In Carruthers and Boucher's (1998a) terminology this is a kind of *cognitive* conception of language "which sees language as crucially implicated in human thinking." (p. 1).

The basic idea is a simple one: language (speech) is a tool for making information available to others that they may use to regulate their activities (Reed, 1996). In particular,

---

[163] These neural circuits may be a kind of *internal control parameter* in Keijzer's (1997, 1998) sense.

[164] It may be possible to incorporate Damasio's (1994, 1999) notion of bodily 'as if loops' into such a perspective (see my discussion in chapter 6). Certainly Damasio is one of the foremost advocates of the idea that imagery is the internally initiated use of perceptual and motor structures.

language "can be used to make people aware of prior or upcoming environmental situations ... [that enables] ... a kind of modified and collectivised prospective control" and it enables us to "select, modify and transform ecological information to serve our own purposes and make a new form of information that is not dependent on the immediate environment. Thus language is often used to make others aware of possible, hypothetical, and even fictitious states of affairs". (Reed, 1996, p.156). By utilising particular utterances (spoken symbol structures) a speaker can provide a listener with information about things that they cannot perceive. So a simple kind of time-space distancing can occur by 'passing on a signal' in the place of (perceivable) environmental information. The next step in the argument is to suggest that, with experience with this *social* symbolic regulation of activity, we can come to use language privately and individually to regulate our own actions in an out-of-context fashion. Finally we can argue that the regular use of activity patterns in out-of-context situations can lead to these patterns becoming engaged automatically and habitually in their new environments. Complex advanced cognitive skills that required a symbolic scaffolding to learn and perfect can later be used without the benefit of private speaking or visualising if they are exercised regularly. Some skills however always require jumping out of context, constructing novel combinations of activity patterns, and re-presenting a 'natural interpretation' in non-natural ways. It is these kinds of skills, which might include planning, troubleshooting, dealing with novelty, overcoming habits, complex reasoning, causal analysis, problem-solving, that will always require symbolic mediation (whether silent and private, or out-loud, or in dialogue with others) (Dennett, 1998).

Although this basic idea seems plausible enough, there are thorny problems ahead for anyone who attempts to flesh it out in more detail.

Problems for an External Symbol View

The first problem, that Wozniak (1976) calls the *language and thought paradox*, is that it is hard to see how it is that you can tell yourself something that you do not already know. What more can one get from saying something to oneself (especially silently) that cannot be got from simply thinking it? Numerous suggestions have been made here: Dennett (1991) suggests that speech may supply a vital connection between two (or more) neural circuits that are neurobiologically unconnected. Language can serve as a 'cognitive autostimulant' that follows an external pathway. Carruthers (1996, 1998) has developed a similar but more detailed theory along these lines. In essence he argues that our linguistic systems are the neurocognitive Grand Central Station of the brain and that the coordination of other modular activities is carried out using the lingua franca of natural language. Language, that is, provides the vehicle for thought that requires the coordination of many evolutionarily basic modular systems (see also Clark, 1997, 1998b).

But for the dedicated anti-representational interactionist there exists a deeper problem because, unlike the previous authors, the non-representationalist suggests that there is *no* prelinguistic thought of a self-directed and context-independent nature available to initiate 'off-line' action (Carruthers, for instance, has no problem with off-line cognizing if it can be plausibly thought to be available in a domain specific way *within* particular modules). The recruitment hypothesis that I am arguing for here requires that, in advanced cognition, 'language comes first' and 'language provides us with the resources for context-independence'. The only kind of presymbolic 'thought' available for the recruitment hypothesis is that occasioned by direct perception and action within the local environment[165]. So my problem is this: how can symbol use occur without advanced thought being already in place? More specifically:

1. How can a self-regulating symbol-user initiate an out-of-context 'train of thought', composed of complex efforts at creating virtual environments, that can be used to produce non-situationally-determined action? That is, how can a situationally-determined basic cognizer suddenly acquire the ability to *autocue*, *self-remind*, or *voluntarily access action patterns*, as Donald (1991, 1998, 1999) puts it?
2. Secondly, how can a 'system' with no explicit, declarative memory (an implicit, procedural, and basic cognitive system) use symbols if it cannot somehow 'store' and 'access' that symbolic knowledge from somewhere inside its head? That is, how can public, external representations do their thing without them being outer expressions of more fundamental internal representations?

I have no fully worked out answer to the first question, but my basic intuition is that what we call *autocuing* or *voluntary recall* is still a kind of situationally-determined cognitive activity[166]; that is, 'voluntary' access to action patterns via a symbolic network is ultimately initiated by things happening in the real environment. Our ability to rely upon symbolic 'virtual environments' provides the opportunity to be determined by a larger environment than that 'lived in' by basic cognitive creatures, but we remain determined nonetheless. This 'real environment autonomy' may at least make advanced cognizers

---

[165] And perhaps the fact that the bodily mechanisms that enable this basic cognition may occasionally operate spontaneously in the absence of the sort of environmental situation they are 'fitted' to, as, for instance, may happen in dreaming.

[166] Despite Donald's (1991, 1998, 1999) argument that the advent of mimetic culture requires the evolution of autocuing or voluntary recall, he nowhere advances an argument about how his theory avoids determinism. Cognitivism is equally susceptible to these 'free-will versus determinism' problems (see Copeland, 1993, chap. 7; Juarrero, 1999).

seem to be more flexible and 'voluntary' than their basic cousins. Deacon (1997) has a similar suspicion that he expresses this way:

> Symbolic analysis is the basis for a remarkable new level of self-determination that human beings alone have stumbled upon. The ability to use virtual reference to build up elaborate internal models of possible futures, and to hold these complex visions in mind with the force of the mnemonic glue of symbolic inference and descriptive shorthands, gives us unprecedented capacity to generate independent adaptive behaviors. ... [S]ymbolically mediated compulsions to act are far more chaotic [than nonsymbolic indexical compulsions], in the technical sense of that word, far more susceptible to the influence of tiny initial differences in starting assumptions or ways of dividing up experiences and qualities symbolically. ... It is not so much that our actions arise from a totally unconstrained and compulsion-free center of intentions, but that the potential starting point, the intended purpose we have modeled, can be drawn from such a vast variety of alternatives with little initial difference in motive power. (p. 434)

Adult human 'free will' is an important issue for any theory of cognition and it should not be ignored. However, to date there do not exist any convincing explanations of how free will can be reconciled with a theory of cognition that is consistent with what is known in the natural sciences.

The second question, of how symbols can extend our cognitive reach beyond the here and now without the symbolic information being explicitly 'encoded' inside the agent, is, I believe, more readily explainable than the first. Such an explanation needs to embrace, what Bechtel (1996) calls, an *external symbol approach* or a *distributed understanding of language*.

A Distributed View of Language

A number of interactionist thinkers have been attracted to the idea that language and other symbol systems might be fruitfully thought of as external structures that mediate our basic, pattern mapping, cognitive abilities (e.g., Bechtel, 1993a, 1996, 1997; Clark, 1997, 1998b; Dennett, 1991, 1996; Rowlands, 1999). The general idea is that *external structures* such as utterances, gestures, diagrams, instruction lists, and plans can be thought of as environmental transformations (effected by our motor skills) that can be responded to with our basic, situated perception-action abilities. One popular example comes from the work of the connectionist researchers Rumelhart, Smolensky, McClelland, and Hinton (1986). They suggest that symbol-use requires three things: 1) an external arena where physical symbols are instantiated, 2) a capacity to comprehend and use those symbols (pattern recognition and association), and 3) a capacity to produce new symbol structures (pattern completion). They illustrate their idea by imagining how a connectionist system could be built to multiply large numbers. Because of the limitations of reasonably sized neural networks, it would probably be impossible to train one to multiply any two large numbers together, but it may be possible to build a pair of interacting networks that can deal with the problem in a human-like fashion. People generally multiply large numbers using place-

value multiplication with a pen and paper (unless they have a calculator handy)[167]. These external scaffolds do not just make the task more efficient or reliable; they actually completely transform the nature of the problem by turning it into a sequence of repeated situated actions. The process starts by giving the problem a physical form by writing it down. We then apply a pattern recognition strategy to locate the first single digit problem. We solve the problem (pattern completion) and then write down the solution to it in the appropriate place in the environment. This transforms our perceived environment and we loop back to our first strategy of search for patterns (the next single digit problem). This problem is solved via pattern completion, and so on. So such a system could be implemented using two connectionist networks (one a pattern recogniser, the other a pattern completer) and a system for manipulating structures in the environment (some kind of 'writer') (Rowlands, 1999, pp. 163-164).

The key lesson of this thought experiment is that it illustrates how a complex compositional problem can be decomposed into the deployment of a series of skilful pattern-mapping procedures (know-how) that are hooked into a modifiable environment. Rumelhart et al. (1986) describe the benefits of this kind of process in the following way:

> Each cycle of this operation involves first creating a representation through manipulation of the environment, then a processing of the (actual physical) representation by means of our well-tuned perceptual apparatus leading to further modification of this representation. By doing this we reduce a very abstract conceptual problem to a series of operations that are very concrete and at which we can become very good. Now this applies not only to solving multiplication problems. It applies as well to solving problems in logic (e.g., syllogisms), problems in science, engineering, etc. These dual skills of manipulating the environment and processing the environment we have created allow us to reduce very complex problems to a series of very simple ones. This ability allows us to deal with problems that are otherwise impossible. This is *real* symbol processing and, we are beginning to think, the primary symbol processing that we are able to do. Indeed, on this view, the external environment becomes an extension to our mind. (pp. 45-46)

The agent-environment system *as a whole* knows how to multiply large numbers, but the entire task does not go on 'in the head'. Rather it consists of the interplay of internal and external structures. People, and the imaginary connectionist system, use paper and pen to 'store' the results of each operation and ultimately produce an answer which we can then read off of the external environmental scaffold. The answer never appears fully in the dynamics of the network, as it were. There is no need to postulate an internal system that deals in mathematical symbols and their transformations. Rather the environment and the 'agent-system' modulate each other's activity until the answer appears inscribed on the

---

[167] At least modern Western people seem to multiply this way. People from different cultural groups use different mathematical methods, such as using an abacus. Moreover, it seems that when people solve mathematical problems mentally they make use of internalised versions of these methods (Hatano, 1982).

environmental surface. Some may object that such a suggestion does not explain our ability to, at least sometimes, do large number multiplication 'in our heads'. But Rumelhart et al. (1986) have a distinctly Vygotskian response to this objection:

> There is one more piece to the story. This is the tricky part and, we think, the part that fools us. Not only can we manipulate the physical environment and then process it, we can also learn to internalize the representations we create, "imagine" them, and then process these imagined representations – just as if they were external (p. 46)

Bechtel (1993a, 1993b, 1996, 1997; Bechtel & Abrahamsen, 1991) and Rowlands (1999) have attempted to generalise Rumelhart et al.'s idea to the much more complex domain of human language. There are three main tasks that face a theorist interested in developing a distributed view of language. First, language needs to be portrayed as having a strong external and environmental flavour. This is not an intuitively obvious way of thinking of language since traditional cognitivist (e.g., Chomskian) views of language take it to be largely an in-the-head affair. Under the distributed view however linguistic symbols are understood to be environmental entities that can be perceived and that are compositional in nature. Words take on a variety of *physical* forms as sound patterns, manual signs, or written characters. These physical forms can be used compositionally because they can be produced in a large variety of syntactic combinations that express composite meanings. Spoken words and manual gestures can be strung together in temporal sequences and written words can be concatenated spatially. Physical symbol tokens are both *generative* or *productive* (they can be combined with, juxtaposed with, and slotted into, other symbol structures to create an unlimited variety of meaningful sequences) and exhibit *systematicity* or *generality* (roughly, by rearranging tokens one can represent different semantic relationships between the same symbolised concepts). Compositionality is understood as embedded in the languages themselves rather than being viewed, as the Chomskian's would have it, as a set of internal neurobiological constraints. So humans have the capacity to *understand* and *use* this compositional feature of linguistic external structures rather than these structures being systems that cause language to be compositional because of their compositional internal structure[168]. Bechtel (1993a) notes that "[t]he key to the external symbol approach is to move formal symbols, which adhere to syntactic rules, out of the head and locate them in the environment of the system." (p. 130). Similarly,

---

[168] See Deacon (1997, chap. 3) for a discussion of how we can think of languages as semi-autonomous external systems that co-evolve with human neural structures and cognitive capacities rather than as simple behavioural expressions of the functioning an internal biological subsystem (such as that envisioned by Chomsky, 1980, 1995). Hutchins and Hazlehurst (1991, 1995; Hazlehurst & Hutchins, 1998) have shown that a simple kind of grammar emerges within a simulated community of interacting connectionist 'agents'. Steels (1998) has attempted a similar demonstration with real robots.

Rowlands (1999) argues that rather than postulating "an internal linguistic or quasi-linguistic system" where language is understood to be *plugged* into the individual's brain, "we plug the organism into a linguistically structured environment, and we give it the capacity to use, to utilize, this environment." (Rowlands, 1999, pp. 179-180).

The second aspect of a distributed approach to language involves taking the enforcement of grammatical principles to be something that is primarily carried out by the language community. "Ultimately, what determines the grammatical correctness or otherwise of combining one symbol with another in a particular way is the pattern of grammatical usage exhibited by the linguistic community." (Rowlands, 1999, p. 178). This stands in opposition to traditional symbolic accounts that argue that the well-formedness of linguistic productions occurs because the internal symbolic system's structure does not allow badly formed language strings. So on the distributed reading, grammatical acceptability is in some sense *learnt*[169]. This means that the underlying neural system need not instantiate a linguistically (compositionally) structured system but, rather, it need only be capable of coming to learn to respect those grammatical relationships that exist in the communicative environment. In other words, language comprehension is understood as an ability to use the connectionist-like skills of pattern-mapping (recognition, association, and completion) to manipulate and use external environmental entities (words, sentences, propositions). Such an internal system is not usefully understood as building an internal model of the linguistic propositions one hears or reads, but of having the procedural resources for skilfully *perceiving* and *using* environmental entities. As Bechtel (1993a) notes "[w]hat exists inside the cognitive system is not an internal representation of these external symbols, but an ability to extract information from them and to produce symbol strings which adhere to the syntactical rules that characterize properly formed strings ..." (p. 130) and that "[t]he network *knows how* to extract information from grammatically structured sentences, but in order to do this it does not have to have an internal representation of the sentence upon which computations can be performed." (p. 138). So connectionist systems might be able to emulate such feats even though they do not possess compositional architectures. Rowlands (1999) argues that "connectionist systems might be

---

[169] This does not mean that simple induction from examples of speech in the environment will do the job. As Chomsky (1959) has famously argued, a simple general learning system could not possibly lead to the kind of grammatical sophistication that humans, even young children, display. Evidence shows that children tend to make a large number of correct guesses when it comes to learning grammar (see Deacon, 1997, chap. 3). However, this does not mean that some sort of learning system (perhaps with built-in learning biases of some kind) cannot learn how to use and generate grammatically correct symbol structures without the benefit of negative examples and explicit feedback about correctness (see, e.g., Christiansen & Chater, 1999; Christiansen, Chater, & Seidenberg, 1999; Deacon, 1997; Redington & Chater, 1998; Seidenberg, 1997).

capable of exhibiting such features as generativity, systematicity, and inferential coherence, not because they themselves are structured in such a way as to exhibit those features, but, rather, because, in virtue of the structure they do have, they are capable of using or employing an external linguistic system which exhibits these features." (p. 181).

The third key aspect of a distributed approach to language is that we need to think of the *production*, as well as the comprehension, of language as involving pattern-mapping processes and environmental manipulation. That is, speech production should be viewed primarily as a motor activity – a bodily skill for creating the appropriate sound patterns, manual gestures, or inscriptions, for getting ideas across, naming things, indicating, issuing orders, regulating the behaviour of others, and so on. Our ability to do this no doubt arises from some perhaps 'innate' abilities for producing sounds such as babbling and cooing (see Pinker, 1994) and innate abilities to chew and swallow (MacNeillage, 1998; Whitcombe, 1996) (what we might call the *intrinsic dynamics* of our oral-vocal-respiratory systems in Goldfield's [1995] and Saltzman's [1995] terms). These abilities are tuned and modulated by practice and environmental and social feedback (the most important being whether appropriate responses to specific communicatory intentions can be extracted from other people and whether one's own sound production sounds right to one's own ear).

This picture of language use as emerging from the interaction of a pair of procedural, pattern-mapping systems (a 'comprehender' and a 'producer') and a highly structured sociolinguistic environment may seem a little too vague for many cognitive scientists' likings. However, there does exist some interesting simulation research that shows that a collection of non-compositional 'neurocognitive innards' may well be able to exhibit compositional skills when trained in an appropriately structured 'linguistic' environment.

Implementing a Distributed Language System

Both Bechtel (1993a, 1996) and Rowlands (1999) examine some recent connectionist efforts for getting networks to treat language-like strings in a compositional manner without a compositionally structured computational architecture (e.g., Elman, 1990, 1993, 1995; St. John & McClelland, 1990). These simulations use recurrent architectures and the sentences to be analysed are fed into the input layer one word at a time. The results of these simulations are intriguing and indicate that a distributed view of language is quite plausible. Specifically, these networks can, to a limited degree, extract ('recognise') the syntactic and semantic/thematic roles that words have in sentences and can produce output that respects such 'interpretations' "given nothing more than a corpus of positive examples of allowable texts – exactly what the UG [universal grammar] theorists had said was impossible" (Deacon, 1997, p. 134). For instance, in Elman's (1993) simulation the network's task was to 'guess' the next word in a sentence that was presented to the network one word at a time. The trained network was quite successful at picking the correct *kind* of

word (i.e., whether it was syntactically appropriate) for sentences that it had already been exposed to in training and also for novel sentences. The network could even cope with embedded clauses and structural transformations. Significantly such success did *not* occur when the network was trained using full, complex sentences right from the outset. In this condition the network rapidly converged on suboptimal solutions that exposure to subsequent learning trials could not overcome. By exposing the network to complex cases from the outset, it was faced with "a very large area of logical space to search for a solution and, therefore, caused it to search wildly for solutions to problems where these solutions, in fact, depended upon solutions to more simple problems." (Rowlands, 1999, p. 185; see also Deacon, 1997, pp. 132-135). In simple terms, artificial neural networks work best when they deal with a training set that contains a large number of relatively similar items. The learning process quickly finds the appropriate approximate input-output mappings and then gradually refines them so that narrower distinctions can be made. However, when faced with a set of quite varied training items, the network cannot quickly find an approximately correct mapping solution because the diverse training mappings possess a variety of quite different solutions. The result of combining a complex training set with an incremental learning strategy, such as backpropagation of error, is that the network rapidly fixes on a suboptimal ('in-between') solution from which it cannot escape.

Elman (1993) found that there were two solutions to this problem. The first solution was to *phase the training*. Basically this boils down to training the network with simple sentences first and, once it had learnt to provide appropriate output for these sentences, to advance to more complex sentences. Rowlands (1999) argues that there is an important sense in which training is phased when children learn language. He suggests that children are gradually exposed to different kinds of language as they mature. At about six months of age infants seem to elicit *infant-directed speech*, a syntactically, semantically, and lexically simplified form of language, from their caregivers. Prior to this point it is not used. Throughout early development adults seem to modulate their language to suit the capacities of the child. However, this in itself is unlikely to be enough, as the real world of linguistic stimulation is no doubt messy, with the child being exposed to complex adult language as well as tailored simple language. So the infant also needs some ability to focus only on the appropriate 'training examples'. Elman's (1993) second method for training his 'language comprehension network' provides an insight into how this might be possible. The second strategy was to *phase the memory*. This basically involved exposing the network to fully complex sentences and depriving it of the feedback it would have normally gained from the context units for later words in each training sentence. Roughly speaking Elman made the network 'ignore' input after it had processed three or four words of a sentence. Over the training session Elman gradually increased the number of words that the network would 'pay attention to' until feedback from the previous input was supplied for all of the words

in each sentence. Rowlands (1999) thinks that such a finding may tell us something about the way a child might need to function in order to learn language:

> Elman's employment of a phased-memory strategy seems to indicate that limitations on certain types of memory might, in fact, be a crucial and positive factor in determining the ability of statistical inference engines to penetrate theoretical domains that are hierarchically ordered in the sense of being formed by way of recursive combinatory rules. (Rowlands, 1999, p. 189)

Deacon (1997) argues that children's language learning seems to ride on just these sorts of 'limitations'. In fact he goes so far as to argue that languages are entities that have evolved to fit the cognitive profile of the developing child:

> Being unable to remember the details of specific word associations, being slow to map words to objects that tend to co-occur in the same context, remembering only the most global structure-function relationships of utterances, and finding it difficult to hold more than a few words of an utterance in short-term memory at a time may all be advantages for language learning. (p. 135)

In sum, then, several recent connectionist simulations provide some intriguing demonstrations that language-like tasks may be implementable using non-compositional systems that are plugged into highly structured linguistic environments. Moreover, their limitations, and the solutions used to overcome them, seem to map roughly onto aspects of the language learning situation.

From Indexical Pattern Mapping to Symbolic Pattern Mapping

All of this talk of connectionist networks exists in the somewhat rarefied atmosphere of computer simulations. Bechtel (1996) suggests that the strategy used by Elman and others may not be the best one for making sense of a distributed account of language. In particular, he argues that these models fall prey to many of the criticisms of 'disembodied connectionist simulations' that were mentioned in the previous chapter. In fact Bechtel (1996) expresses some very 'interactionist' concerns:

> I am not even convinced that that they represent the most fruitful way of exploring linguistic ability within a connectionist or nonsymbolic framework. They approach the task of language processing in isolation from other cognitive activities and the needs of the organism to control its body in its environment. Thus, there is no real semantics for such models. A far more realistic approach might begin with a model of a system functioning in an environment, that is, a system with sensory capacities to absorb information from its environment and motor capacities to change its environment. (pp. 74-75)

Deacon's (1996, 1997) theory of language evolution focuses closely on semantic issues and provides some ideas for extending the connectionist-inspired notion of a distributed language system. Deacon suggests that pattern associating connectionist networks can be understood to be indexical systems – systems that can learn to associate a token (or, to use the terminology of classical conditioning, an unconditioned stimulus) with a referent (or conditioned stimulus). Thus, something like a pattern-associating ability may underpin the ability of many animals to learn how to modulate their behaviour (their response) when

exposed to the token. Deacon argues that all non-human animal communication involves an ability to deal with indexical reference. That is, animals modulate the behaviour of others by producing calls, gestures, and displays that are associated with particular referents. Vervet monkeys, for instance, learn that 'barking' is associated with the presence of a predator such as a leopard that requires running up a tree, that 'coughing' is associated with eagles and safety can be found by hiding in shrubs, and that 'chuttering' is associated with the presence of a snake and getting together into a group to crowd about it in a threatening manner[170] (Cheney & Seyfarth, 1990). In other words, pattern-associative learning enables the modulation of behaviour with respect to distant environments when a conspecific produces a sign (call, gesture, etc.)[171]. In sophisticated animals (many mammals, birds such as parrots) it may even possible to use an index (say a word spoken by a human trainer) in a variety of contexts via stimulus generalisation and the transfer of learning sets (see Deacon, 1997, pp. 80-81). This may give the appearance of a kind of symbolic competence because it appears that the animal appreciates the meaning of the index rather than just knowing what goes with what.

However, Deacon argues that indexical, pattern-associating abilities differ from symbolic abilities in two important ways. First, indexical reference requires an animal to *observe* a correlation between the index and its referent. Moreover, if the experience of correlation ceases the index will eventually cease to 'mean' the referent. This phenomenon can be observed in the extinction phases of operant conditioning experiments. Symbols, on the other hand, still mean their referent, even if we hardly ever, or never, observe a correlation between the presence of the symbol token and its referent (after we have learnt what the symbol refers to). Somehow, once symbolic reference is created, it is 'maintained' without the benefit of an indexical relationship to a referent.

The second difference between symbols and indices is that learning an indexical relationship does not affect other known indexical associations. By contrast, when we learn a new symbol (e.g., a new word) it becomes importantly associated with other words (symbolic referential relationships) that we also know. Deacon argues that the former difference is explained by the latter:

> We do not lose the indexical associations of words, despite a lack of correlation with physical referents, because the possibility of this link is maintained implicitly in the stable associations between words. It is by

---

[170] Actually, it is not at all clear whether the referents of the vervets' calls are particular animals, kinds of animals, or signals for behaving in a certain way (see Cheney & Seyfarth, 1990; Noble & Davidson, 1996).

[171] This is roughly what Christensen and Hooker (2000, in press) mean by *constructive* (i.e., modifiable, learnable) gradient tracking.

virtue of this sort of dual reference, to objects and to other words ..., that a word conveys the information necessary to pick out objects of reference. ...

This referential relationship between the words – words systematically indicating other words – forms a system of higher-order relationships that allows words to be *about* indexical relationships, and not just indices in themselves. But this is also why words need to be in context with other words, in phrases and sentences, in order to have any determinate reference. Their indexical power is *distributed*, so to speak, in the relationships between words. Symbolic reference derives from *combinatorial* possibilities and impossibilities, and we therefore depend on combinations both to discover it (during learning) and to make use of it (during communication). (Deacon, 1997, pp. 82-83)

In essence Deacon argues that in learning to 'deal with' (produce and comprehend) symbolic reference, people come to grasp the relations that exist among the symbol tokens themselves and not just those that hold between the tokens and objects. Tokens have all kinds of semantic and syntactic relationships with each other. Syntactically some token combinations are allowed and some are not. Semantically we might learn that the tokens **cat** and **dog** are quite closely associated compared to, say, **cat** and **bus**. This means that humans have the capacity to detect and use *combinatorial patterns* – that is, patterns made up of rule-based combinations of stimuli (symbol tokens), and not just holistic patterns. Deacon expresses this idea by saying that people come to make sense of the high-order regularities that hold among tokens. He argues that this is a hard problem for nonhumans because it involves moving from the indexical referential strategy of associating objects and tokens to the symbolic referential strategy that downplays object-token association in favour of token-token association. However, due to our oversized prefrontal cortex, Deacon believes that we humans are well predisposed to this kind of learning in a way that other animals are not.

The prefrontal cortex helps us inhibit the tendency to act on simple correlative stimulus relationships and guides our sampling of alternative higher-order sequential or hierarchic associations. Its role in language and symbol learning in particular is not, however, merely to increase something we might call prefrontal intelligence. Rather I suspect the importance of the size change can be thought of in displacement terms, in patterns of cognition as in patterns of brain development. Prefrontal computations out-compete other cognitive computations and tend to dominate learning in us as in no other species. In simple terms, we have become predisposed to use this one cognitive tool whenever an opportunity presents itself, because an inordinate amount of control of the other processes in the brain has become vested in our prefrontal cortex. The way the parietal cortex handles tactile and movement information, the way the auditory cortex handles sound information, the way the visual cortex handles visual information are all now much more constrained by prefrontal activity than in other species. (Deacon, 1997, p. 265)

In other words, Deacon argues that human beings have brains that are well-suited to learning about symbolic relationships as well as indexical relationships. Mature human brains 'contain' *symbolic pattern mappers*. This does not mean that humans have anything like a subpersonal physical symbol system in their heads. Rather it means that humans are specially sensitive to networks of token-token relationships as well as to token-referent

relationships. In an environment that contains symbolic systems – systems of tokens (spoken words, manual gestures, text) that are internally (grammatically) organised as well as externally (referentially) organised – humans are capable of grounding reference courtesy of an ability to detect token-token patterns. Whereas the kind of time-space distancing enabled by indexical pattern mapping is limited to observed covarying relationships, an ability to appreciate symbolic patterns enables an agent to introduce new tokens by relating them to tokens that already exist in the network. This means that reference can be 'grounded' without having to observe token-referent relationships. Moreover, it also enables one to create or use tokens that have no real-world referents but that gain their meanings purely from their relationships with other symbol tokens. Specifically,

> [b]ecause of our symbolic abilities, we humans have access to a novel higher-order representation system that not only recodes experiences and guides the formation of skills and habits, but also provides a means of representing features of the world that no other creature experiences, the world of the abstract. We do not just live our lives in the physical world and our immediate social group, but also in a world of rules of conduct, beliefs about our histories, and hopes and fears about imagined futures. (Deacon, 1997, p. 423)

Deacon (1997) goes on to suggest that such abilities derive from the reuse of basic cognitive mechanisms.

> This world [the world of the abstract] is governed by principles different from any that have selected for neural circuit design in the past eons. We possess no brain regions specially adapted for handling the immense flood of experiences from this world, only those adapted for life in a concrete world of percepts and actions. (p. 423)

These ideas tie in nicely with the earlier discussion of advanced cognition as an ability to image – that is, the ability to use our basic situated perception-action skills in an off-line manner.

Symbols and Concepts in a Sociocultural Milieu

So far I have been rather vague about the nature of the relations that are supposed to hold between tokens in a symbolic system and what the consequences of these relationships are for symbol users. Fortunately there exists a long tradition within sociolinguistics, social theory, and some parts of psychology that examines the interactions that hold between symbol systems, social structure, and cognition. For instance, Lock and Symes (1996) argue that the evidence from sociolinguistic, psychological, and anthropological research

> ... broadly supports a view of communicative symbol systems as the providers of 'cognitive technologies or 'tool-kits' that variously afford analytic, context-independent thought. This is not to claim that language determines thought, but that conceptual systems can make some universal mental operations more or less difficult to carry out. As an example of what we mean by this, we draw an analogy with the differences between the Roman and Arabic number systems. The mental operations of multiplication and division are available to users of both systems, but the symbolic 'tool-kits' affect the ease with which these operations

can be performed: it is easier to divide 63 by 9 to get 7 than LXIII by IX to get VII. Similarly, it is easier to think about, and even perceive, the structure of society and how communication systems operate within it if one possesses a symbol system which symbolizes these things. (p. 230)

So symbol systems, especially languages, are not just concept-neutral tools for expressing ideas, as the communicative theorists would have it, but they are also things that re-present the world in a particular way. They are, in essence, cultural conceptual systems, so in learning to use these symbol systems, people learn about these conceptual systems. Concepts, in this sense, are cultural entities that correspond to a collection of skills that 'belong together' in a particular community's activities. Hendriks-Jansen (1996), borrowing extensively from Clark (1993), puts it this way:

Concepts do not correspond to specific internal states or structures. They must be equated with the ability to exercise a particular skill: the skill of deploying the concept in all relevant situations that do not lead to category mistakes. Internally, such a skill may be sustained by a variety of disjunct computational states that need to have nothing in common from the point of view of physics, neuroscience, or even computational psychology. They are held together purely by the fact that, together, they enable an agent to "negotiate some macro-level domain which interests us in virtue of our daily human life." (p. 284)

Thus, to use an analogy from Clark (1993), being able to use a symbol, that stands for a concept like **dog**, is like being a skilled golfer. The skills that hang together as golf skills, such as hitting balls over long distances, then over smaller distances into little holes, and so on, are culturally, not neurobiologically or computationally related. Similarly

being good at thinking about dogs and acting appropriately with respect to dogs (which is what having the concept of "dog" amounts to) does not mean being good at manipulating some internal symbol that can be identified with the word "dog" or the various properties of dogs. The different skills involved in having thoughts about dogs do not share a common causal factor that may be found within the person who has mastered the concept. These skills are only *about* the concept because they converge on the public linguistic and cultural entity "dog," in the same way as the rearing skills of the mother duck are *about* her duckling only because they converged on the duckling "in the real world." (Hendriks-Jansen, 1996, p. 284)

This perspective shows how concepts, as external public entities, can be understood as a collection of linked situated, on-line skills. In learning to grasp communal concepts one is effectively learning how a variety of perception-action skills 'go together'. Some may be skills that are directly relevant to the immediate situation and thus are the kind of skills that a fully basic cognizer might use. But many others may normally (for a basic cognizer) relate to 'distant environmental situations'. In the symbolic/conceptual framework of the agent's community these skills are considered to be relevant to the current situation. It is through the use of symbol systems (primarily language) that agents gain access to these relationships.

It is entirely possible that the neural/bodily mechanisms that enable the diverse skills that underpin a concept become neurally 'connected', perhaps via reentrant signalling or some other kind of constructive neurobiological process. But this should not be viewed as

evidence in favour of the cognitivist strategy of trying to explain cognitive activity purely in terms of neurobiological phenomena. This would be a bit like trying to understand body building purely in terms of muscle physiology while ignoring key factors such as the biomechanical effects of using weight resistance and the properties of training programmes. The 'glue' that holds together the situated perception-action skills exists initially only in the set of semantic relationships that exist within communal symbolic/conceptual networks (i.e., languages and 'discourses').

Is there any evidence that learning to use language enables an ability to cognize in an advanced fashion – to be able to 'think' in an abstract and off-line manner? I believe that a good number of the findings in early cognitive developmental psychology are consistent with this hypothesis.

The Guided Reinvention of Advanced Cognition

Mature human beings do not simply live in their immediate environments. Perceived objects and events remind us of other things and, in off-line cogitating, we can consider things which seem to have nothing at all to do with our current environmental situation. How is it that humans become able to do such things? By and large the evidence suggests that we do not begin life with such an ability. As newborns we enter the world as basic, situated cognizers for whom the immediate environment is fairly much all there is to life. Many theories of cognitive development propose, sometimes only implicitly, that, over the first few years of life, we develop some kind of internal time-space distancing mechanism. But, from an interactionist perspective, such a move does not seem to be workable. A promising alternative, as I have been suggesting in this chapter, is to argue that immersion in a human social environment enables children to grasp ways in which the here-and-now may relate to distant environments (e.g., Lock, 1980, 1991). It is worth investigating whether the dynamics of a human environment can help construct a child's developing time-space distancing abilities. Within some of the child development literature there is suggestive evidence that communicative interactions between infants and others provides the requisite social scaffolding for eliciting time-space stretching skills.

For instance, Nelson (1996) points out that around the middle of the child's first year

> Parents tend to institute signals, explicit and implicit, that provide guidance to the child's anticipation and participation. For example, a mother may routinely ask, "Do you want lunch now?" before heading to the kitchen, or, "How about a bath?" as the accustomed hour draws near. Babies toward the end of the first year respond to these verbal, and nonverbal, signs, typically following or even preceding the parent in the expected direction. (p. 97)

Parents seem to be providing symbolic links between events in the here and now with 'distant' events. Lock (1991; Lock et al., 1989) refers to this phenomenon whereby an adult (or older peer or sibling) makes explicit the 'distant' implications of things in the

here-and-now, as *context hopping*. In context hopping situations the child is encouraged to expand what they 'see' when they perceive events and objects. Language is being used by the senior partner to attach objects and events from distant environments to things in the local environment. Lock (1991) gives the following examples from observations he and his colleagues (Lock, Service, Brito, & Chandler, 1989) have made of an 18 month old interacting with his mother.

Example 1. "It's like the clock at home." Here she is showing the child a clockwork toy, and focusing his attention on the key that winds it up.

Example 2. "It's a snake. Remember the one you had at the fairground once?" Now she is showing him a different toy, but still relating what he sees to another temporally displaced context.

Example 3. "He's got boots on, hasn't he? They look like Daddy's shoes." At this point they are looking at a picture book, and again she betrays her presupposition that her infant understands a lot of what she is saying, that her words serve to bring to his mind, as the picture does in hers, a remembered event.

Example 4. Still looking at the book, she says: "Ooh, it's a red car, isn't it, like the one we saw in the garage when we got petrol this morning." (p. 289)

Lock (1991) describes what he believes are the effects of these sorts of interactions on the child in the following way:

Unless asked to remember, [the infant] might not: He might stay context bound. In all the aforementioned examples, left to his own devices, the infant might not be motivated to exercise what we call "his memory," to see (literally) this event in front of him as having any connection to an event that can be accessed only through the directive and depictive content of his mother's words (and, later, it is claimed that the mother's strategy not only is acting in a constitutive manner with respect to her child's memory skills, but is changing, inter alia, his experience of objects and events as he perceives them). (p. 289)

During this period, between about 15 and 24 months, children begin to engage in, what Lock (1993) calls, *symbolic referential communication*. Prior to the second half of the second year children use words in a holistic manner that is usually strongly tied to the immediate context "such that when a child says "down," for example, as it puts an object down, it means all of <I am putting that object down>, and is not coding just one item of meaning." (Lock, 1993, p. 284). However towards the end of the second year children seem to come to grasp the insight that individual objects have names. At this point a 'naming explosion' or 'vocabulary burst' occurs where there is a sudden increase in the vocabulary learning rate. What seems to be happening here is that the child is moving from learning words in, what Deacon (1996, 1997) refers to as, an *indexical manner* (the kind of learning accomplished by association of a sound or gesture and an object or event) to learning words in a *symbolic manner* where there exists a "generalizable "understanding" of the relation between sounds and objects, a relation that holds over differing objects and contexts." (Lock, 1993, p. 283). Thus children come to use words to refer to *elements* of a

situation and not the entire situation[172]. They have moved from using words holistically to using them predicatively:

> The separate components coded in the adult language [begin to be] constructed: actions, in English, need to be constructed as separate from agents and patients: where "kick" might originally mean <I am kicking the ball>, it has come to just represent the act of kicking; similarly for "ball" uttered in the same circumstances. Once these differentiations are made, then the child is moving itself beyond the realm of reference to that of predication; not just identifying something as a ball, but identifying a ball and saying something about it. (Lock, 1993, p. 284)

A number of authors have suggested that the advent of an ability to *predicate* signals the beginning of an ability to think beyond the current context using symbolic tools because the predicate can refer to an out-of-context (unperceivable) feature of the subject (e.g., Hendriks-Jansen, 1996; Lock, 1993; Reed, 1996; Russell, 1996). At this stage words are beginning to be lifted out of their embedded context and have become at least slightly 'decontextualised'[173]. Words are not *just* used to accompany or move along events as they have been up until this point. However, this is still the primary function of language at this age (Nelson, 1996). Once the child has mastered an ability to comment, they have the *potential* to explicitly link something in the immediate environment with something beyond the current context. Typically this pre-grammatical predication is not enough for making any complex statements about things beyond the immediate context. Yet pre-grammatical predication does provide a communicative environment where adults and older siblings and peers can help young children extract the implications of their speech and make more explicit the meanings of the words they use. And it is in this environment that we see children and caregivers engaging in context-hopping 'dialogues'. Nelson (1996) and others have argued that it is these sorts of social interactions which provide the necessary scaffolding for coming to grips with more powerful grammatically-based symbol structures.

Between the ages of about two and four years children enter an *opening to language* phase where they can actively incorporate what is said to them into their understanding of the world (Nelson, 1996). At this time children acquire much of the grammar of their community's language(s) (Lock, 1993). In grammatically sophisticated language many of the aspects of communicated information are explicitly marked by words, inflections,

---

[172] Lock et al. (1989) have observed that at around this age many parents direct their child's attention to objects, parts, and relations that hold between them when, for instance, looking at picture books. They suggest that this might provide an opportunity for coming to grips with the idea that words map onto elements of a scene.

[173] In terms of Karmiloff-Smith's (1992, 1994) approach, discussed in the previous chapter, children are making their knowledge more *explicit*.

word-order and so on. So at this stage of development, children have learnt that symbols can be combined to produce very specific meanings that do not depend upon there being a shared context between the speaker and the listener (Lock, 1999; Lock & Symes, 1996). They have, in the terms of Deacon's theory, learnt how to ground reference by appreciating a large number of token-token relationships (both syntactic and semantic). This radically expands the child's ability to use symbols to go beyond the here and now. When one can string together a complex symbol structure all kinds of relationships to distant, abstract, and imaginary objects and events can be made with some element of the local environment. However, between two and four years this ability is still far from adult-like. Children of this age can both understand another person's utterances and can use it to modify their own world knowledge, yet, as noted in the previous section on cognitive development, they cannot easily appreciate that what they are told may constitute a different point of view from their own. Nelson (1996) claims that these children cannot "maintain two simultaneous mental representations of the same situation ..." (p. 129). Linguistic information, it seems, cannot be held apart as representing someone's point of view as distinct from one's own. It either feeds in to one's own view of the world or it is ignored.

At about four or five years of age children overcome this limitation. Nelson (1996) calls this the *language-modelling* phase. In this phase another's discourse can be used by the individual to construct an independent alternative point of view that can be juxtaposed with one's own point of view. It is quite possible that specific sophisticated grammatical developments enable this sort of ability. For instance, de Villiers and de Villiers (2000) have shown that children's ability to appropriately understand *complements* (e.g., She thought/said/claimed that the cake was in the cupboard) predicts success at false belief tasks (but not vice versa). To understand a complement a listener must be able appreciate that a proposition can be true even though part of it may be false (e.g., 'Tom thinks that the world is flat' may be true even though 'the world is flat' is false). That is, de Villiers and de Villiers' research suggests that it is developments in particular linguistic skills which underpin success in cognitive tasks that involve appreciating virtual environments and not the other way around.

It appears that about the age of four, most children have developed a near-adult ability to comprehend and produce symbolic patterns that weld together aspects of the immediate context with distant, and even abstract, objects and events. Moreover, they can consider these symbolic relations independently of their own personal understanding of the world. This is truly an ability to *imagine* in the sense that most adults mean it.

Conclusion

How does a distributed view of language help us to understand advanced cognizing from an interactionist perspective? Crudely speaking, the idea pursued here is that external public symbols can serve as 'stimuli' for triggering an appropriate activity pattern in the same way that the symbol's referent might. "There's a mugger around the corner" can initiate mugger-avoiding behaviour in the same sort of way as the sight of someone lurking suspiciously nearby. A cognitivist might argue that advanced cognitive power of language and other symbol systems is parasitic on the deeper pre-existing cognitive power of thought which is understood as a computational process involving subpersonal representations. By contrast, the distributed view of language suggests that it is the external symbol system itself, which is concretely realised by actual language usage within a community, that can 'connect' things in the current environment with things that are temporally or spatially distant or are abstract in nature. Language-users begin life as possessors of a basic cognitive package of situated perception-action skills that enable them to learn to use the symbol system to, in a sense, access the world beyond the concrete here-and-now. No subpersonal internal representation system is required in this scenario only an ability to respect and use the relationships that hold between the symbols that make up an external symbol system.

In language production a person perceives something in their local environment and *knows how* to produce an appropriate linguistic response. In language comprehension a person perceives an external symbol structure and *knows how* to produce an appropriate behavioural response which itself may be the production of another external symbol structure. In both cases knowing how can arise from some kind of pattern-mapping device that does not need to possess a compositional architecture. Some sort of sophisticated connectionist system may do the job. In skilled self-directed speech, where the same agent plays the role of both producer and comprehender, these same coupled pattern-mapping processes can go on with a minimal amount of 'efferent leakage'. This may *look* as if some kind of subpersonal internal symbolic computing is going on, but it is in fact a case of internalization or the private appropriation of personal-level social skills. This echoes Vygotsky's (1930-1935/1978) claim that "[e]ach function in the child's cultural development appears twice: first, on the social level, and later, on the individual level; first, between people (interpsychological), and then inside the child (intrapsychological)." (p. 57).

This interactionist view of advanced cognition may seem to be a little too behaviourist for many cognitive scientists' tastes. For many, such a position might seem to be an attempt to resurrect Skinner's (1957) approach to language that Chomsky (1959) so successfully demolished. However the interactionist position must be much more sophisticated than Skinner's. It will require a detailed theory of the 'cognizer's innards' that, perhaps, builds

on research with embodied neural networks. It will also require a much more complex understanding of the environment than can be formulated in terms of operants, reinforcers, and punishers. And, perhaps most importantly, it will require a sophisticated understanding of the interaction dynamics that exist in the development, learning, and practicing of symbolic skills. Such a theory will be something like a sociolinguistic extension of Christensen and Hooker's (2000, in press) notion of gradient tracking. From this perspective symbol structures form a special kind of environmental gradient; environment regularities that can be used to expand one's interactive horizon. Moreover, symbol structures are gradients that humans can *produce* in almost any situation and that are constantly skilfully modified. Christensen and Hooker argue that intelligent behaviour occurs in systems that are capable of self-directed anticipative learning. These are, roughly, systems that can tune and even construct new gradient tracking processes. Learning to use language can be usefully understood in these terms. In production we are constantly learning how to best use language to provide a 'gradient' that effectively regulates the behaviour of others (or oneself in the case of self-directed speech). In comprehension we are constantly modifying our gradient tracking skills in trying to make the best use of the symbol structures produced by others[174].

In an important sense, the view I have developed here does not quite mesh with the simple claim that I started with. Initially I suggested that advanced cognition involves the 'activation' of situated activity patterns by external symbols. Perhaps a better way to understand the influence of external symbol systems on cognition is in terms of their ability to expand an agent's *Umwelt* or cognitive horizon (e.g., Lock, 1999). This broadening of the cognitive horizon is much larger and more varied than one that can be had by an 'indexical' cognizer because symbol systems enable an agent to both 'associate' objects and events that they have never *seen* together, and associate objects and events with imaginary, hypothetical, and abstract entities. An indexical cognizer could perhaps associate a token for referent$_1$ with referent$_2$ but the result would not be an association of referent$_1$ and referent$_2$. Rather the token would either become associated with both referents (i.e., it would have a disjunctive content) or, given time, the association with referent$_1$ would be extinguished. A symbol-user, on the other hand, might, for example, come to associate a place with death or happiness simply because they had been told about such an association. Experience with a symbolic environment enables one to 'see' much more in the local environment than one would from a basic cognitive point of view. One can, sometimes slowly and deliberately, spin out a 'symbolic gradient' that can help modulate

---

[174] This does not just include a pure semantic appreciation, but also more importantly includes skill in detecting pragmatic and paralinguistic subtleties such as irony, lying, and honesty.

one's behaviour with respect not just to simple causal regularities, but also with respect to possibilities, likelihood's, desired goals, and fears. A skill with symbols fundamentally transforms the way the world looks and feels.

9. Where is Cognition?

Like Humpty Dumpty, brain, body, and world are going to take a whole lot *of putting back together again.*

Andy Clark (1997, p. 222).

The traditional way of replying to inquiries about the whereabouts of cognition is to point to a location in the head, specifically to the brain. This, I believe, is the primary commitment of cognitivism. The representational and computational hypotheses central to modern cognitivist theory derive from an attempt to understand how time-space distancing can be achieved by systems that 'cognize in their heads'. The mental representation system provides the critical subpersonal mechanism for 'stretching' the influence of the environment over space and time. Many cognitivists argue that creative, compositional, representational systems, such as those possessed by humans, enable the even more spectacular cognitive skill of modulating behaviour with regard to possible, hypothetical, counterfactual, and imaginary states of affairs.

The intrinsic internalism apparent in these cognitivist assumptions can be understood as the result of a long scientific and philosophical history of attempts to locate the mechanism necessary for the spontaneity, animacy, and sentience of living things, especially of animals and even more especially of human beings, within part of the body. Descartes has often been criticised by modern theorists for a lack of mechanistic zeal – for preserving the truly human in some mysterious non-materialistic *esprit*. Yet he also clearly articulated the position that most animals are mere machines explainable by natural philosophy. Cognitivism reveals its ties to Descartes in two important ways. First, it has attempted to scientifically validate Descartes' idea that there is a single locus of cognition within the body – the brain has become the instructor, the controller, and the thinker, for the entire person. Second, cognitivism has relegated bodily (non-brain) processes, reflexes, and sensorimotor activity to the realm of the 'merely mechanical'; the important stuff goes on in the central neural circuits.

Forces within the sciences have been struggling with this Cartesian and Newtonian picture of the merely mechanical for the entire twentieth century. In recent decades, with the advent of fields such as non-linear thermodynamics, it has become obvious that the world of matter does not just consist of linear, passive, equilibrium, reversible reactions (Prigogine & Stengers, 1985; Swenson & Turvey, 1991). The picture of the billiard-ball universe has been shattered. Scientists are beginning to appreciate how complex, environmentally sensitive, dynamic order can arise quite naturally in even relatively simple physical systems. A growing number of scientists argue that these natural dynamic processes may also be at work in living systems. These ideas are beginning to filter into explanations of animal and human behaviour in psychology and the cognitive sciences. Incorporating these influences into psychology is no small task. Even simple animals such

as insects are much more complicated than self-organising physical systems like the Belousov-Zhabotinsky reaction. It is more than likely that the current tools used for analysing these relatively simple physical dynamical systems will be woefully inadequate for understanding people and other cognizers. Despite this probability, current dynamical systems strategies and methods already provide an opportunity for challenging many of the assumptions of the cognitivist computational approach to cognition.

Embodied, situated, and distributed interactionism has arisen within this intellectual climate where explanations of complex order ride on an appreciation of issues such as decentralisation, emergence, complex cooperativity, and constraint satisfaction (Resnick, 1994). In this thesis I have attempted to tie together some of the insights of researchers who are open-minded about the prospects of moving beyond the current representational-computational framework by understanding behaviours, including human behaviours, in distributed interactionist terms. This final chapter summarises the framework sketched thus far and offers some comments about the ways in which interactionist theoretical and empirical work might be developed and encouraged.

The Lessons of Interactionism

Agent-Environment Systems and Autonomy

In chapter 4 I enumerated several themes that recur in many of the embodied, situated, and distributed interactionist fields of research that were examined in chapter 3. Perhaps the key theme is that cognition should be understood as something that occurs within an agent-environment system. On the surface this may seem to be a rather trite claim, but it in fact signals the biggest difference between interactionism and cognitivism. Of course, cognitivism has always allowed the environment a role in cognition. In that tradition the environment provides input and a place for motor output. However, the concept of the agent-environment system goes, or should go, much further than this. What this system concept implies is that there is no such *place* as the mind. Rather, there exists minded activity which involves a complex interaction between neural, bodily, and environment events. In this view the world does not need to be understood as getting into people's heads so that thoughtful behaviour can be produced. In interactionism environmental entities stay in the environment and the agent modulates their behaviour by detecting and manipulating those entities. People, and other cognizers, are viewed as complex systems who are sensitive to the environment and who use the environment in their day-to-day projects which, at root, involve maintaining their autonomy. In making reference to the notion of autonomy, interactionism commits itself to a vision of the cognizer as inherently concerned with staying alive by making sure that they act appropriately in the environment. Autonomy implies that cognizers are, at base, systems that possess biological norms for

maintaining their systemic integrity. Cognizers are fundamentally systems that sense, coordinate with, modulate, and manipulate the world.

Interactionist Innards

Our understanding of the role of the cognizer's neural and bodily mechanisms changes when we conceive of the cognizer as primarily designed to detect and manipulate environmental entities. In cognitivism the mind/brain is viewed as an information assembler and user – a small biological factory for building interpretations, ideas, and plans. An interactionist detect-and-manipulate architecture, on the other hand, is best understood in terms of a regulatory, coordinative, and integrative nervous system coupled with a complex body. Functionally speaking, this system consists of perceptual instruments and performance instruments that are wired together by a complex of neural control structures. These instruments are, perhaps, wired together in a kind of subsumption architecture made up of semi-autonomous, behaviour-layers that have high-bandwidth connectivity within a layer and low-bandwidth external connectivity among the layers. These are essentially complex sensorimotor control systems. On this view, behaviour is not understood as existing preformed in some kind of neurally realised plan or template. Rather, behavioural activity is understood to emerge from subtle, homeostatic couplings between the topography of the local environment and the supple dynamic activity of sensory systems and action systems. Imagine a water-filled balloon being balanced on a single finger. The balloon constantly wobbles as its centre of gravity changes and the finger-hand-arm system continuously feels for and compensates for these wobbles by way of a continuous sequence of tiny motor movements. These movements then alter the dynamics of the balloon and the centre of gravity changes again requiring more finger-hand-arm movements. Think of the balloon as the environment and of the finger system as the agent. This is the interactionist's guiding image.

Complex, cultivated activity, to use Hughlings Jackson's term (Whitcombe, 1996), emerges from the fact that these behaviour systems can be modified by learning. Many animals are capable of tuning and assembling their basic sensorimotor control systems in the light of experience, practice, and play (Christensen & Hooker, 2000, in press). From an internal, subpersonal perspective, control parameters are explored and associative connections are forged or modified. From a whole-animal perspective new and improved skills for exploratory and performatory action are learned.

Basic Cognitive Abilities

In may seem puzzling as to how these situated, sensorimotor systems manage to modulate their behaviour when the sensed environmental entities do not make contact with the agent's body. Most theorists claim that this is not a problem because distal perception

ensures that the agent and the relevant environmental entity can remain in contact. Often this relationship is theorised in terms of some kind of information transfer between the environment and the agent's sensory surfaces. This relationship is analogous to the way a message is passed down a phone line by a caller to a receiver. This image is no doubt at work in the use of communication science's information theory by psychologists and other cognitive scientists. This image is probably responsible for the idea that the cognizer must actively (computationally) mine the sparse raw data given in the message from the environment for an 'intended' or original message (see Knapp, 1986). These ideas lie at the heart of the cognitivist understanding of perception and force on us a representational understanding of even quite basic cognitive relationships between animals and their environments.

However, distal perception can be understood in another way in terms of contact with the 'causal consequences' of the environmental entity on the local environment (e.g., Thomas, 1999). In this essentially Gibsonian view, the local environmental arrays can be understood to be composed of potentially trackable gradients that the organism can follow or avoid (Christensen & Hooker, 2000, in press). In Gibson's (1979/1986) terminology gradients *specify* the existence of particular environmental object and events. Animals possess sensory systems that are sensitive to these gradients. An animal's *Umwelt* is constituted by these specifications of 'distant' objects and events. However, animals do not just wait about or randomly chance upon these gradients. Instead, they actively and skilfully explore and interrogate their surroundings in search of them. Animals manipulate their environments in two ways (Rowlands, 1999): they carry out exploratory, or epistemic, actions for getting environmental information and performatory, or pragmatic actions in order to serve their material needs.

Thus, a distal-perceptual gradient tracking explanation promises to keep the interactionist image intact in the face of the problem of coordination with distant, but still causally potent, environmental entities. The water-filled balloon analogy suggested earlier can be modified by thinking of the water-filled balloon as being balanced on a stick that is grasped by the hand-arm system. This is a more complex coupled system but it remains a detect-and-manipulate system. No message is transmitted from balloon to hand, yet agent and environment remain coupled.

Advanced Cognitive Abilities

The interactionist vision can potentially explain a good number of the time-space distancing abilities of humans and other animals. I have argued that a basic, situated cognitive creature realised by an autonomous system with a regulatory, coordinative, and integrative nervous system can modulate its behaviour with regard to current distal environmental entities, future states of the environment that are specified in the current

local environmental structure (anticipation), and past states of the environment (basic, recognition-type memory and pattern-mapping learning). But what about the, perhaps uniquely human, abilities for thinking about things that are truly out of context, in the sense that there exists nothing at all in the local environment that can be used as an index of their presence? These are genuinely representation hungry situations. Surely here the notion of a mental representation must come into its own (Clark, 1997; Clark & Toribio, 1994). Representation hungry situations are a deep problem for the determined interactionist. The subtle detect-and-manipulate powers of situated robots built with subsumption architectures or controlled by embodied neural networks seem far from capable of exercising such advanced cognitive powers.

The suggestion I have made is that a shared external symbol system might be able to transform the situated cognizer into a representation-savvy creature without us needing to postulate the existence of an internal mental modelling subsystem. Many animals coordinate their behaviour with respect to things in the environment that they cannot directly sense via the communicatory efforts of others of their kind. These social animals produce calls or displays that can inform others of the presence of particular environmental objects or events. This kind of indexical ability to use calls and displays as referential signs (Deacon, 1996, 1997) can be understood as a kind of social/communal gradient-tracking phenomenon where the communicatory behaviour of conspecifics serve as gradients. Humans have developed very sophisticated varieties of these social-communicatory gradient systems. We call them languages.

Languages instantiate shared conceptual systems and they exist concretely in terms of interlocking systems of narratives and discourses. Roughly speaking, these are publicly shared systems that specify the relations that hold between environmental entities and the kinds of behaviours that are considered appropriate responses to those entities. Because languages specify relationships that hold between in-context entities and out-of-context entities they can be used to access those out-of-context entities. Once the agent has access to such entities, they can serve as 'stimuli' for the modulation of behaviour. Thus, in this view, language constitutes a kind of portable environmental gradient production (and comprehension) system for flexibly regulating behaviour. People learn to use languages but their compositional syntactic and semantic structure is not realised in the organisation of their neural architecture. Rather, various neural/bodily systems provide the pattern-mapping mechanisms necessary for skilfully producing and comprehending utterances in appropriate circumstances (Bechtel, 1993a, 1996; Rowlands, 1999).

This is admittedly an incomplete theory of advanced cognition, but it is one that I believe is worth pursuing. In order to elaborate this framework, it will be necessary to study language and its cognitive consequences in social and ecological settings and to pay close

attention to the ways in which syntactic, semantic, and pragmatic aspects of language-use work together to realise language's powerful regulatory and modulatory properties. Such research is already being carried out in the fields of artificial life (Hutchins & Hazlehurst, 1991, 1995; Parisi, 1997) and situated robotics (Steels, 1998). This research complements existing traditions in psychology and other social and cognitive sciences that seek to understand the role of language and narrative in the 'social construction' of human cognitive skills (e.g., Harré & Gillett, 1994; Lock, 1980, 1999; Lock & Symes, 1996; Nelson, 1996). The theoretical positions of these different researchers may not always be in agreement, but I believe that they are sufficiently similar in their aims for there to be some fertile exchanges of ideas that could develop and bolster the interactionist vision of advanced cognition that I have described in this thesis.

Studying Cognition Interactionist Style

Little of what I have discussed in the previous chapters makes direct contact with the phenomena examined, and terminology utilised, within traditional cognitive psychology. Thus, there is a risk that interactionism will evolve into an obscure side discipline "specializing in slime molds and six-legged insects ..." (Hendriks-Jansen, 1996, p. 1) that is ignored by cognitive psychologists and their cognitive science kin. Unfortunately, there is no simple way to provide alternative interactionist explanations of such things as attention, explicit memory, working memory, and problem-solving. Such tasks are formidable ones that will require a lot of time and much empirical research and theoretical analysis. However, what can be done at this stage is to outline the kind of strategies that an interactionist might employ for making sense of the various cognitive phenomena that feature centrally in modern cognitive psychological research.

A reasonable number of the research findings I have described have been discovered using traditional experimental techniques in laboratory settings. But a good deal of interactionist-inspired research has arisen because researchers have embraced a variety of non-traditional (at least within psychology) procedures and strategies for studying cognitive activity. In order to investigate issues such as perception-action cycles, environmental manipulation, and interactive emergence researchers will need to alter or supplement existing methods of research. Broadly speaking, there seem to be three kinds of strategy that will assist the interactionist: an emphasis on detailed observation, an experimental methodology that preserves important agent-environment relations, and an experimental rather than purely analytic approach to modelling.

Detailed Observation

Much of the interactionist-oriented empirical research that I have discussed in the previous chapters involves giving close attention to the fine detail of the actual activity that people

and other cognizers engage in. For instance, Ballard et al.'s (1995, 1997) research involved the careful monitoring of eye and hand movements, and Hutchins (1995a, 1995b) spent many days observing and recording the activities of people at work in various natural environments. Hendriks-Jansen (1996) notes that two of the key orienting attitudes of ethology are "start all analysis from a secure base of description" and "view behavior in the context of the environment to which it has been adapted." (p. 205). He argues that cognition researchers should also embrace these methodological assumptions. Detailed descriptive research is facilitated when researchers can easily access instruments, such as video cameras, for the careful recording of microgenetic interactions. Hendriks-Jansen observes that the explosion in microgenetic infant developmental research began when such instruments became available in the mid 1970s. Detailed observation of cognitive activity in natural environments makes it possible to grasp the kinds of small-scale temporally-extended interactions that occur between agents and their environments (which may include other agents). The ability to gather such information is obviously important for interactionist research. It also enables theorists to understand cognitive activity in dynamical systems terms because such research can throw up thickly described time series data sequences. Perhaps the most important reason for spending a good deal of time engaged in such exploratory research is that the true complexity of real-world cognitive activity can be appreciated with such a strategy.

Experiments that Preserve Interaction

Once a researcher has a good understanding of how cognizers act in their natural environments, they can explore the ways in which varying environmental conditions affect the production of cognitive activity in more controlled, experimental conditions. Experiments, broadly understood, involve the systematic and controlled manipulation of the experimental situation. This enables the researcher to elicit information about the powers and limitations of cognizers that cannot be gleaned from simple observation. From an interactionist perspective, it is important that many of the essential features of real-world cognitive activity are preserved in experimental research. For instance, ecological psychologists often vary the environmental conditions in their experiments in a similar way to most traditional psychological researchers. However, unlike many of their colleagues, ecological experimenters often arrange their experiments so that participants can actively control and manipulate their environments. The ecological psychologist's interest is in seeing how people modulate their exploratory and performatory actions in response to various kinds of environmental information. Such research manages to preserve the perception-action cyclic nature of cognition within the experimental situation. This stands in contrast to many traditional experiments where what the participant perceives is completely controlled by the experimenter and the actions of the participant cannot

influence the environmental dynamics. The coupling relationship between perception and action is effectively severed. Thus, well-designed experimental research built upon a solid descriptive base can illuminate an interactionist understanding of cognition.

Experimental Modelling

The third useful strategy for engaging in interactionist research is the building of simulated cognizing systems such as mobile robots. This use of simulation need not have the goal of building artificial cognizers. Rather, it should be understood as a computer-assisted attempt at constructing theoretical models of cognitive activity. Traditionally scientists have either formulated models in some natural or mathematical language or have built physical models of their target phenomena. Modern computers radically expand the scope and power of modelling possibilities by enabling theorists to construct complex, dynamic, and rigorous simulations of their target phenomena. Within behavioural research robotics offers a method of combining these powerful computational simulations with physical models. Of course, classical cognitivist researchers also build 'cognitive simulations' but the approach taken differs from that used by situated robotics and artificial life researchers. In classical simulation work much of the designing is planned prior to the building of the system by using formal task descriptions to map out relevant computational-level algorithms. By contrast, situated roboticists and artificial life researchers utilise a 'situated action strategy' where a simple system is constructed and then tried out in an environment to see how it copes. Information gleaned from such investigation is used to debug and tweak the design in order to make it function more effectively. The process continues by adding new components to the system or by altering its environment to see how such changes affect its activity. Gradually, through an ongoing process of trial and error a useful, functional model can be developed. This sort of simulation research is more like perfecting a recipe for baking a cake than the classical approach, which is akin to the drawing up of a building blueprint by an architect. The advantage of such experimental modelling is that it explicitly incorporates a 'reality-testing procedure' where real-world constraints, such as noise and slippage, bring to light important and useful lessons about engineering the internal dynamics of the simulated agent. For instance, Horswill (1992) notes that what I have called the *experimental modelling approach* can lead to fortuitous discoveries such as learning that a pair of crude sensor systems can compensate for each other's inadequacies, or, that a subsystem designed to enable one kind of activity actually solves an unrelated problem.

These kinds of research strategies promise to focus the light of scientific inquiry more strongly upon those often overlooked aspects of human and animal activity that interactionist's believe are deeply implicated in the production of adaptive and intelligent behaviour. Interactionism is, at its core, a position that takes the study of cognition to be a

fundamentally multidisciplinary task that moves beyond the study of what goes on in people's heads. Modern cognitive science has achieved a lot through its interdisciplinary alliances, but its focus has remained the investigation of the individual and their brain. Interactionism's emphasis on the importance of the non-neural body, the topography of the physical environment, and social and cultural factors in cognitive activity, suggests that many other disciplines could be fruitfully incorporated into this cooperative enterprise. Currently cognitive science has very little to say about the ways in which people and animals use and relate to their surroundings, but environmental psychologists, geographers, ecologists, ergonomics and human factors researchers, and landscape researchers have tackled these issues from a variety of perspectives. Medical researchers, kinestheologists, and biologists of many stripes have detailed appreciations of the way bodies work. Sociologists, ethnographers, cultural anthropologists, and sociolinguists can tell us a good deal about the nature of social groups and cultures. If cognition is all about how animals, particularly we human animals, 'know how to do things', and cognitive science involves trying to understand this knowing, surely we should tap into the deep reservoirs of knowledge and expertise embodied in and distributed among all of these research traditions.

References

Abraham, F. D., Abraham, R., & Shaw, C. D. (1991). *A visual introduction to dynamical systems theory for psychology*. Santa Cruz, CA: Aerial Press.

Adolph, K. E., Eppler, M. A., & Gibson, E. J. (1993). Crawling versus walking: Infant's perception of affordances for locomotion on slopes. *Child Development, 64,* 1158-1174.

Agre, P. E. (1993). The symbolic worldview: Reply to Vera and Simon. *Cognitive Science, 17,* 61-69.

Agre, P. E. & Chapman, D. (1987). Pengi: An implementation of a theory of activity. *Proceedings of the Sixth National Conference on Artificial Intelligence*, 196-201.

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K. & Watson, J. D. (1989). *Molecular biology of the cell* (2nd ed.). New York: Garland.

Allard, T., Clark, S. A., Jenkins, W. M., & Merzenich, M. M. (1991). Reorganization of somatosensory area 3b representations in adult owl monkeys after digital syndactyly. *Journal of Neurophysiology, 66,* 1048-1058.

Almassy, N., Edelman, G. M., & Spoorns, O. (1998). Behavioral constraints in the development of neuronal properties: A cortical model embedded in a real-world device. *Cerebral-Cortex, 8,* 346-361.

Anderson, J. A. (1991). On what building a Martian three-wheeled iguana tells us about complex minds. *Behavior and Philosophy, 19,* 91-102.

Anderson, J. R. (2000). *Cognitive psychology and its implications* (5th ed.). New York: Worth.

Annett, J. (1995). Motor imagery: Perception or action. *Neuropsychologia, 33,* 1395-1417.

Armstrong, D. M. (1968). *A materialist theory of the mind.* London: Routledge.

Ashby, W. R. (1960). *Design for a brain* (2nd ed.). New York: Wiley.

Baars, B. J. (1986). *The cognitive revolution in psychology.* New York: Guilford Press.

Baillargeon, R. (1986). Representing the existence and the location of hidden objects: Object permanence in 6- and 8-month-old infants. *Cognition, 23,* 21-41.

Baillargeon, R. (1987). Object permanence in 3.5- and 4.5-month old infants. *Developmental Psychology, 23,* 655-664.

Baillargeon, R. (1993). The object concept revisited: New directions in the investigation of infants' physical knowledge. In C. E. Granrud (Ed.), *Visual perception and cognition in infancy* (pp. 265-316). Hillsdale, NJ: Erlbaum.

Baillargeon, R., Spelke, E. S., & Wasserman, S. (1985). Object permanence in five-month-old infants. *Cognition, 20,* 191-208.

Ballard, D. (1991). Animate vision. *Artificial Intelligence, 48,* 57-86.

Ballard, D. H., Hayhoe, M. M., & Pelz, J. B. (1995). Memory representations in natural tasks. *Journal of Cognitive Neuroscience, 7,* 66-80.

Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rajesh, P. N. R. (1997). Deictic codes for embodiment in cognition. *Behavioral and Brain Sciences, 20,* 723-767.

Barkow, J. H., Cosmides, L., & Tooby, J. (Eds.). (1992). *The adapted mind: Evolutionary psychology and the generation of culture.* New York: Oxford University Press.

Barlow, G. W. (1968). Ethological units of behaviour. In D. Ingle (Ed.), *The central nervous system and fish behaviour* (pp. 217-232). Chicago: University of Chicago Press.

Barlow, R. B. (1990, April). What the brain tells the eye. *Scientific American, 262,* 66-71.

Barton, D. & Hamilton, M. (1996). Social and cognitive factors in the historical elaboration of writing. In A. Lock & C. R. Peters (Eds.), *Handbook of human symbolic evolution* (pp. 793-858). Oxford: Clarendon Press.

Bates, E., Thal, D., & Janowsky, J. S. (1992). Early language development and its neural correlates. In S. J. Segalowitz & I. Rapin (Eds.), *Handbook of neuropsychology* (Vol. 7, pp. 68-110). Amsterdam: Elsevier.

Bechara, A., Damasio, A. R., Damasio, H., & Anderson, S. (1994). Insensitivity to future consequences following damage to human prefrontal cortex. *Cognition, 50,* 7-12.

Bechtel, W. (1988). *Philosophy of mind: An overview for cognitive science.* New Jersey: Lawrence Erlbaum Associates.

Bechtel, W. (1993a). The case for connectionism. *Philosophical Studies, 71,* 119-154.

Bechtel, W. (1993b). The path beyond first-order connectionism. *Mind and Language, 8,* 531-539.

Bechtel, W. (1996). What knowledge must be in the head in order to acquire language? In B. M. Velichkovsky & D. M. Rumbaugh (Eds.), *Communicating meaning: the evolution and development of language* (pp. 45-78). Mahwah, NJ: Lawrence Erlbaum.

Bechtel, W. (1997). Embodied connectionism. In D. M. Johnson & C. E. Erneling (Eds.), *The future of the cognitive revolution* (pp. 187-208). New York: Oxford University Press.

Bechtel, W. (1998). Representations and cognitive explanations: Assessing the dynamicist's challenge in cognitive science. *Cognitive Science, 22,* 295-318.

Bechtel, W. & Abrahamsen, A. (1991). *Connectionism and the mind.* Oxford: Blackwell.

Bechtel, W. & Richardson, R. C. (1993). *Discovering complexity: Decomposition and localization as strategies in scientific research.* Princeton, NJ: Princeton University Press.

Beer, R. D. (1995a). A dynamical systems perspective on agent-environment interaction. *Artificial Intelligence, 72,* 173-215.

Beer, R. D. (1995b). Computational and dynamical languages for autonomous agents. In R. E. Port & T. van Gelder (Eds.), *Mind as motion: Explorations in the dynamics of cognition* (pp. 121-147). Cambridge, MA: MIT Press.

Beer, R.D. (2000). Dynamical approaches to cognitive science. *Trends in the Cognitive Sciences, 4,* 91-99.

Beer, R. D. & Gallagher, J. (1992). Evolving dynamical neural networks for adaptive behavior. *Adaptive Behavior, 1,* 91-122.

Bem, S. & Keijzer, F. (1996). Recent changes in the concept of cognition. *Theory & Psychology, 6,* 449-469.

Berk, L. E. (1986). Relationship of elementary school children's private speech to behavioral accompaniment to task, attention, and task performance. *Developmental Psychology, 22,* 671-680.

Berk, L. E. (1992). Children's private speech: An overview of theory and the status of research. In R. M. Diaz & L. E. Berk (Eds.), *Private speech: From social interaction to self-regulation* (pp. 17-53). Hillsdale, NJ: Lawrence Erlbaum.

Berk, L. E. (1994, November). Why children talk to themselves. *Scientific American, 271,* 60-65.

Berk, L. E. & Potts, M. K. (1991). Developmental and functional significance of private speech among attention-deficit hyperactivity disordered and normal boys. *Journal of Abnormal Child Psychology, 19,* 357-377.

Bivens, J. A. & Berk, L. E. (1990). A longitudinal study of the development of elementary school children's private speech. *Merrill Palmer Quarterly, 36,* 443-463.

Block, N. (1978). Troubles with functionalism. In C. W. Savage (Ed.), *Perception and cognition: Issues in the foundations of psychology. Minnesota studies in the philosophy of science.* (Vol. 9., pp. 261-325). Minneapolis, MN: University of Minnesota Press.

Block, N. (1986). Advertisement for a semantics for psychology. *Midwest Studies in Philosophy, 10,* 615-678.

Boden, M. A. (1996a). Autonomy and artificiality. In M. Boden (Ed.), *The philosophy of artificial life* (pp. 95-108). Oxford: Oxford University Press.

Boden, M. A. (Ed.). (1996b). *The philosophy of artificial life.* Oxford: Oxford University Press.

Braddon-Mitchell, D. & Jackson, F. (1996). *Philosophy of mind and cognition.* Oxford: Blackwell.

Brooks, R. A. (1991a). Intelligence without representation. *Artificial Intelligence, 47,* 139-159.

Brooks, R. A. (1991b, September 13). New approaches to robotics. *Science, 253,* 1227-1232.

Brooks, R.A. (1997). From earwigs to humans. *Robotics and Autonomous Systems, 20,* 291-304.

Brooks, R. A., Breazeal, C., Irie, R., Kemp, C. C., Marjanovic, M., Scassellati, B., & Williamson, M. M. (1998). Alternative essences of intelligence. *Proceedings of the 15th national conference on artificial intelligence,* 961-968.

Brooks, R. A. & Stein, L. A. (1994). Building brains for bodies. *Autonomous Robots, 1,* 7-25.

Bruner, J. S., Greenfield, P., & Olver, R. (1966). *Studies in cognitive growth.* Cambridge, MA: Harvard University Press.

Brunswick, E. (1952). *The conceptual framework of psychology.* Chicago: University of Chicago Press.

Burton, G. (1993) Non-neural extensions of haptic sensitivity. *Ecological Psychology, 5,* 105-124.

Buss, D. M. (1995). Evolutionary psychology: A new paradigm for psychological science. *Psychological Inquiry, 6,* 1-30.

Capra, F. (1998). *The web of life: A new synthesis of mind and matter.* London: Flamingo.

Carruthers, P. (1996). *Language, thought and consciousness: An essay in philosophical psychology.* Cambridge: Cambridge University Press.

Carruthers, P. (1998). Thinking in language? Evolution and modularist possibility. In P. Carruthers & J. Boucher (Eds.), *Language and thought: Interdisciplinary themes* (pp. 94-119). Cambridge: Cambridge University Press.

Carruthers, P. & Boucher, J. (1998). Introduction: opening up options. In P. Carruthers & J. Boucher (Eds.), *Language and thought: Interdisciplinary themes* (pp. 1-18). Cambridge: Cambridge University Press.

Case, R. (1985). *Intellectual development: Birth to adulthood.* London: Academic Press.

Chalmers, D. J. (1990). Mapping part-whole hierarchies into connectionist networks. *Artificial Intelligence, 46,* 47-75.

Chalmers, D. J. (2000). *A computational foundation for the study of cognition.* St Louis, MO: Philosophy/ Neuroscience/ Psychology Research Report, Washington University.

Chemero, A. P. (1998a). *How to be an anti-representationalist.* Unpublished doctoral dissertation, Indiana University.

Chemero, A. P. (1998b). *Representation and "reliable presence".* Paper presented at the conference of the European Society for Philosophy and Psychology in Lisbon, Portugal (August 1998).

Chemero, A. P. (1999). *Anti-representationalism and the dynamical stance.* Unpublished manuscript, Indiana University.

Cheney, D. L. & Seyfarth, R. M. (1990). *How monkeys see the world.* Chicago: University of Chicago Press.

Chomsky, N. (1957). *Syntactic structures.* The Hague: Mouton.

Chomsky, N. (1959). A review of B. F. Skinner's "verbal behavior". *Language, 35,* 26-58.

Chomsky, N. (1980). Rules and representations. *Behavioral and Brain Sciences, 3,* 1-61.

Chomsky, N. (1995). Language and nature. *Mind, 104,* 1-61.

Christensen, W. D. and Hooker, C. A. (2000). An interactivist-constructivist approach to intelligence: Self-directed anticipative learning. *Philosophical Psychology, 13,* 5-45.

Christensen, W. D. and Hooker, C. A. (in press). Organised interactive construction: The nature of autonomy and the emergence of intelligence. *Communication and Cognition.*

Christiansen, M. H. & Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science, 23,* 157-205.

Christiansen, M. H., Chater, N. & Seidenberg, M.S. (Eds.). (1999). Connectionist models of human language processing: Progress and prospects [Special issue]. *Cognitive Science, 23* (4).

Churchland, P. M. (1989). *A neurocomputational perspective: The nature of mind and the structure of science.* Cambridge, MA: MIT Press.

Churchland, P. M. (1995). *The engine of reason, the seat of the soul.* Cambridge, MA: MIT Press.

Churchland, P. S. & Sejnowski, T. J. (1992). *The computational brain.* Cambridge, MA: MIT Press.

Churchland, P. S., Ramachandran, V. S., & Sejnowski, T. J. (1994). A critique of pure vision. In C. Koch & J. L. Davis (Eds.), *Large-scale neuronal theories of the brain* (pp. 23-60). Cambridge, MA: MIT Press.

Clancey, W. J. (1989). The frame of reference problem in cognitive modelling. *Proceedings of the annual conference of the Cognitive Science Society,* 107-114.

Clancey, W. J. (1997). *Situated cognition: On human knowledge and computer representations.* Cambridge: Cambridge University Press.

Clark, A. (1989). *Microcognition: Philosophy, cognitive science and parallel distributed processing.* Cambridge, MA: MIT Press.

Clark, A. (1993). *Associative engines: Connectionism, concepts and representational change.* Cambridge, MA: MIT Press.

Clark, A. (1996a). Happy couplings: Emergence and explanatory interlock. In M. Boden (Ed.). *The philosophy of artificial life* (pp. 262-281). Oxford: Oxford University Press.

Clark, A. (1996b). Review of Cognition in the Wild. *Philosophical Psychology, 9,* 393-395.

Clark, A. (1997). *Being there: putting brain, body, and world together again.* Cambridge, MA: MIT Press.

Clark, A. (1998a). Embodied, situated, and distributed cognition. In W. Bechtel & G. Graham (Eds.), *A companion to cognitive science* (pp. 506-517). Oxford: Blackwell.

Clark, A. (1998b). Magic words: How language augments human computation. In P. Carruthers & J. Boucher (Eds.), *Language and thought: Interdisciplinary themes* (pp. 162-183). Cambridge: Cambridge University Press.

Clark, A. & Chalmers, D. (1995). *The extended mind.* St. Louis, MO: Philosophy/ Neuroscience/ Psychology Research Report. Washington University.

Clark, A. & Grush, R. (1997). *Towards a cognitive robotics.* St Louis, MO: Philosophy/ Neuroscience/ Psychology Research Report, Washington University.

Clark, A. & Karmiloff-Smith, A. (1993). The cognizer's innards: A psychological and philosophical perspective on the development of thought. *Mind and Language, 8,* 487-519.

Clark, A. & Toribio, J. (1994). Doing without representing? *Synthese, 101,* 401-431.

Cliff, D. (1991). Computational neuroethology: a provisional manifesto. In J.-A. Meyer and S. W. Wilson (Eds.), *From animals to animats: proceedings of the first international conference on simulation of adaptive behavior* (pp. 29-39). Cambridge, MA: MIT Press.

Cliff, D., Husbands, P., & Harvey, I. (1993). Evolving visually guided robots. In J.-A. Meyer, H. L. Roitblat, & S. W. Wilson (Eds.), *From animals to animats 2: Proceedings of the Second International Conference on Simulation of Adaptive Behavior* (pp. 374-383). Cambridge, MA: MIT Press.

Cohen, N. J. & Squire, L. R. (1980). Preserved learning and retention of pattern-analyzing skill in amnesia using perceptual learning. *Cortex, 17,* 273-278.

Cole, M. & Engeström, Y. (1993). A cultural-historical approach to distributed cognition. In G. Salomon (Ed.), *Distributed cognitions: psychological and educational implications* (pp. 1-46). Cambridge: Cambridge University Press.

Copeland, B. J. (1993). *Artificial intelligence: A philosophical introduction.* Oxford: Blackwell.

Corballis, M. C. (1991). *The lopsided ape: evolution of the generative mind.* New York: Oxford University Press.

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition, 31,* 187-276.

Cosmides, L, & Tooby, J. (1994). Beyond intuition and instinct blindness: Towards an evolutionarily rigorous cognitive science. *Cognition, 50,* 41-77.

Cosmides, L. & Tooby, J. (1987). From evolution to behavior: Evolutionary psychology as the missing link. In J. Dupré (Ed.), *The latest on the best: Essays on evolution and optimality* (pp. 277-306). Cambridge, MA : MIT Press.

Costall, A. (1995). Socializing affordances. *Theory & Psychology, 5,* 467-482.

Costall, A. & Leudar, I. (1996). Situating action I: Truth in the situation. *Ecological Psychology, 8(2),* 101-110.

Coulter, J. (1979). *The social construction of mind: Studies in ethnomethodology and linguistic philosophy.* London: Macmillan Press.

Coulter, J. (1983). *Rethinking cognitive theory.* London: Macmillan Press.

Coulter, J. (1989). *Mind in action.* Cambridge: Polity Press.

Crane, T. (1995). *The mechanical mind: A philosophical introduction to minds, machines and mental representation.* London: Penguin.

Craver-Lemley, C., Arterberry, M. E., & Reeves, A. (1997). Effects of imagery on vernier acuity under conditions of induced depth. *Journal of Experimental Psychology: Human Perception and Performance, 23,* 3-13.

Craver-Lemley, C. & Reeves, A. (1987). Visual imagery selectively reduces vernier acuity. *Perception, 16,* 599-614.

Crick, F. H. C. & Asanuma, C. (1986). Certain aspects of the anatomy and physiology of the cerebral cortex. In J. L. McClelland, D. E. Rumelhart, & The PDP Research Group (Eds.), *Parallel distributing processing, Volume 2: Psychological and biological models* (pp. 333-371). Cambridge, MA: MIT Press.

Cummins, R. (1989). *Meaning and mental representation.* Cambridge, MA: MIT Press.

Cummins, R. (1996). *Representations, targets, and attitudes.* Cambridge, MA: MIT Press.

Damasio, A. R. (1989). Time-locked multiregional retroactivation: A systems-level proposal for the neural substrates of recall and recognition. *Cognition, 33,* 25-62.

Damasio, A. R. (1994). *Descartes' error: emotion, reason, and the human brain.* New York: Putnam.

Damasio, A. R. (1999). *The feeling of what happens: Body and emotion in the making of consciousness.* New York: Harcourt Brace.

Damasio, A. R. & Damasio, H. C. (1994). Cortical systems for retrieval of concrete knowledge: The convergence zone framework. In C. Koch & J. L. Davis (Eds.), *Large-scale neuronal theories of the brain* (pp. 61-74). Cambridge, MA: MIT Press.

Damasio, A. R., Tranel, D. & Damasio, H. C. (1991). Somatic markers and the guidance of behavior: Theory and preliminary testing. In H. S. Levin & H. M. Eisenberg (Eds.), *Frontal lobe function and dysfunction* (pp. 217-229). New York: Oxford University Press.

Davies, M. (1995). Consciousness and varieties of aboutness. In C. Macdonald & G. Macdonald (Eds.), *Philosophy of psychology: Debates on psychological explanation* (pp. 356-392). Oxford: Blackwell.

Dawkins, R. (1976). *The selfish gene.* Oxford: Oxford University Press.

De Villiers, J. G. & de Villiers, P. A. (2000). Linguistic determinism and the understanding of false beliefs. In P. Mitchell & K. J. Riggs (Eds.), *Children's reasoning and the mind* (pp. 191-228). Hove, England: Psychology Press.

Deacon, T. W. (1996). Prefrontal cortex and symbol learning: Why a brain capable of language evolved only once. In B. M. Velichkovsky & D. M. Rumbaugh (Eds.), *Communicating meaning: the evolution and development of language* (pp. 103-138). Mahwah, NJ: Lawrence Erlbaum

Deacon, T. W. (1997). *The symbolic species: the co-evolution of language and the brain.* New York: W. W. Norton.

Deheane, S. (1998). *The number sense: How the mind creates mathematics.* London: Penguin.

Dennett, D. C. (1978). *Brainstorms: Philosophical essays on mind and psychology.* Cambridge, MA: MIT Press.

Dennett, D. C. (1991). *Consciousness explained.* London: Penguin.

Dennett, D. C. (1993). Learning and labeling. *Mind and Language, 8,* 540-548.

Dennett, D. C. (1995). *Darwin's dangerous idea: Evolution and the meanings of life.* New York: Simon & Schuster.

Dennett, D. C. (1996). Kinds of minds: *Towards an understanding of consciousness.* London: Weidenfeld & Nicolson.

Dennett, D. C. (1998). Reflections on language and mind. In P. Carruthers & J. Boucher (Eds.), *Language and thought: Interdisciplinary themes* (pp. 284-294). Cambridge: Cambridge University Press.

Deposito, M., Detre, J. A., Aguirre, G. K., Stallcup, M., Alsop, D. C., Tippet, L. J., & Farah, M. J. (1997). A functional MRI study of mental image generation. *Neuropsychologia, 35*, 725-730.

Deutsch, G., Bourbon, T. W., Papanicolaou, A. C., & Eisenberg, H. M. (1988). Visuospatial tasks compared via activation of regional cerebral blood flow, *Neuropsychologia, 26*, 445-452.

Dewey, J. (1969). The new psychology. In G. E. Axtelle et al. (Eds.), *The early works of John Dewey* (Vol. 1, pp. 48-60). Carbondale, IL: Southern Illinois University Press. (Original work published 1884).

Diamond, A. (1985). The development of the ability to use recall to guide action, as indicated by infants' performance on AB. *Child Development, 56*, 868-883.

Diamond, A. (1988). Differences between adult and infant cognition: is the crucial variable presence or absence of language? In L. Weiskrantz (Ed.), *Thought without language* (pp. 337-370). Oxford: Clarendon Press.

Diaz, R. M. and Berk, L. E. (Eds.). (1992). *Private speech: from social interaction to self-regulation.* Hillsdale, NJ: Lawrence Erlbaum.

Diaz, R. M., Neal, C. J. & Vachio, A. (1991). Maternal teaching in the zone of proximal development: A comparison of low- and high-risk dyads. *Merrill Palmer Quarterly, 37,* 83-107.

Diaz, R. M., Padilla, K. A. & Weathersby, E. K. (1991). The effects of bilingualism on preschoolers' private speech. *Early Childhood Research Quarterly, 6,* 377-393.

Donald, M. (1991). *Origins of the modern mind: three stages in the evolution of culture and cognition.* Cambridge, MA: Harvard University Press.

Donald, M. (1998). Mimesis and the executive suite: Missing links in language evolution. In J. R. Hurford, M. Studdert-Kennedy & C. Knight (Eds.), *Approaches to the evolution of language: Social and cognitive bases* (pp. 44-67). Cambridge: Cambridge University Press.

Donald, M. (1999). Preconditions for the evolution of protolanguages. In M .C. Corballis & S. E. G. Lea (Eds.), *The descent of the mind: psychological perspectives on hominid evolution* (pp. 138-154). Oxford: Oxford University Press.

Dretske, F. I. (1981). *Knowledge and the flow of information.* Cambridge, MA: MIT Press.

Dretske, F. I. (1986). Misrepresentation. In R. J. Bogdan (Ed.), *Belief, content and function* (pp. 17-36). Oxford: Clarendon Press.

Dretske, F. I. (1988). *Explaining behavior.* Cambridge, MA: MIT Press.

Dretske, F. I. (1995). *Naturalizing the mind.* Cambridge, MA: MIT Press.

Dreyfus, H. L. & Dreyfus, S. E. (1986). *Mind over machine: The power of human intuition and expertise in the era of the computer.* New York: Free Press.

Edelman, G. M. (1987). *Neural Darwinism: The theory of neuronal group selection.* New York: Basic Books.

Edelman, G. M. (1992). *Bright air, brilliant fire: On the matter of mind.* New York: Basic Books.

Eimas, P. D. (1985, January). The perception of speech in early infancy. *Scientific American, 252,* 35-40.

Eliasmith, C. (1996). The third contender: A critical examination of the dynamicist theory of cognition. *Philosophical-Psychology, 9,* 441-463.

Eliasmith, C. (2000). *How neurons mean: A neurocomputational theory of representational content.* Unpublished doctoral dissertation, Washington University, St. Louis.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14,* 179-211.

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition, 48,* 71-99.

Elman, J. L. (1995). Language as a dynamical system. In R. F. Port & T. van Gelder (Eds.), *Mind as motion: Explorations in the dynamics of cognition.* (pp. 195-225). Cambridge, MA: MIT Press.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: a connectionist perspective on development.* Cambridge, MA: MIT Press.

Evans, G. (1982). *Varieties of reference.* Oxford: Oxford University Press.

Everson, S. (1991). Introduction. In S. Everson (Ed.), *Psychology: Companions to ancient thought 2* (pp. 1-12). Cambridge: Cambridge University Press.

Eysenck, M. W. (Ed.). (1990). *The Blackwell dictionary of cognitive psychology.* Oxford: Blackwell.

Eysenck, M. W. & Keane, M. T. (1995). *Cognitive psychology: A student's handbook* (2nd ed.). Lawrence Erlbaum Associates: Hove, London, and Hillsdale.

Farah, M. J. (1989). The neuropsychology of mental imagery. In F. Boller & J. Grafman (Eds.), *Handbook of neuropsychology* (Vol. 2, pp. 395-413). Amsterdam: Elsevier.

Farah, M. J. (1994). Neuropsychological inference with an interactive brain: a critique of the "locality" assumption. *Behavioral and Brain Sciences, 17,* 43-104.

Farmer, J. D. (1990). A rosetta stone for connectionism. *Physica D, 42,* 153-187.

Fischer, K. W. and Biddell, T. (1991). Constraining nativist inferences about cognitive capacities. In S. Carey & R. Gelman (Eds.), *The epigenesis of mind: Essays on biology and cognition* (pp. 199-235). Hillsdale NJ: Lawrence Erlbaum.

Fivush, R. & Schwarzmueller, A. (1998). Children remember childhood: Implications for childhood amnesia. *Applied Cognitive Psychology, 12,* 455-473.

Flach, J. (2000). *Research on information form in human-machine interfaces: A meaning processing approach.* (Report for the Japan Atomic Energy Research Institute [JAERI]). Dayton, OH: Wright State University.

Flanagan, O. J. (1991). *The science of mind* (2nd ed.). Cambridge, MA: MIT Press.

Flavell, J. H., Green, F. L., & Flavell, E. R, . (1996). The development of children's knowledge about attentional focus. *Developmental Psychology, 31,* 706-712.

Flavell, J. H., Green, F. L., & Flavell, E. R. (1986). Development of knowledge about the appearance-reality distinction. *Monographs of the Society for Research in Child Development, 51* (1), Serial No. 212.

Flavell, J. H., Green, F. L., & Flavell, E. R., (1995). *Young Children's knowledge about thinking. Monographs of the society for research in Child Development, 60*(1, Serial No. 243).

Flavell, J. H., Green, F. L., & Flavell, E. R.. (1993). Children's understanding of the stream of consciousness. *Child Development, 64,* 387-398.

Flavell, J. H., Green, F. L., Flavell, E. R., & Grossman, J. B. (1997). The development of children's knowledge about inner speech. *Child Development, 68,* 39-47.

Fodor, J. A. (1975). *The language of thought.* New York: Crowell.

Fodor, J. A. (1980). Methodological solipsism considered as a research strategy in cognitive psychology. *Behavioral and Brain Sciences, 3,* 63-109.

Fodor, J. A. (1986). Why paramecia don't have mental representations. *Midwest Studies in Philosophy, 10,* 3-23.

Fodor, J. A. (1987). *Psychosemantics: The problems of meaning in the philosophy of mind.* Cambridge, MA: MIT Press.

Fodor, J. A. (1990). *A theory of content and other essays.* Cambridge, MA: MIT Press.

Fodor, J. A. & Pylyshyn, Z. W. (1981). How direct is visual perception? Some reflections on Gibson's 'ecological approach'. *Cognition, 9,* 139-196.

Fodor, J. A. & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition, 28,* 3-71.

Franceschini, N., Pichon, J.-M., & Blanes, C. (1991). Real time visuomotor control: From flies to robots. *Proceedings of the IEEE 5th International Conference on Advanced Robotics,* 91-95.

Frauenglass, M. H. & Diaz, R. M. (1985) Self-regulatory functions of children's private speech: A critical analysis of recent challenges to Vygotsky's theory. *Developmental Psychology, 21,* 357-364.

Geertz, C. (1983). *Local knowledge: further essays in interpretative anthropology.* New York: Basic Books.

Georgopolous, A. P., Lurito, J. T., Petrides, M., Schwartz, A. B., & Massey, J. T. (1989, January 13). Mental rotation of the neuronal population vector. *Science, 243,* 234-236.

Gibson, J. J. (1966). *The senses considered as perceptual systems.* Boston: Houghton Mifflin.

Gibson, J. J. (1986). *The ecological approach to visual perception.* Boston: Houghton Mifflin. (Original work published 1979).

Gibson, K. R. (1996). The biocultural human brain, seasonal migrations, and the emergence of the upper-Palaeolithic. In P. Mellars & K. Gibson (Eds.), *Modelling the early human mind* (pp. 33-46). Cambridge: The McDonald Institute for Archaeological Research.

Giddens, A. (1984). *The constitution of society.* Berkeley and Los Angeles: University of California Press.

Giere, R. N. (1988). *Explaining science: A cognitive approach.* Chicago and London: University of Chicago Press.

Glenberg, A. M. (1997). What memory is for. *Behavioral and Brain Sciences, 20,* 1-55.

Globus, G. G. (1992). Toward a noncomputational cognitive neuroscience. *Journal of Cognitive Neuroscience, 4,* 299-310.

Godfrey-Smith, P. (1988). Review of Ruth Garrett Millikan's Language, thought, and other biological categories, *Australasian Journal of Philosophy, 66,* 556-560.

Goldenberg, G., Müllbacher, W., & Nowak, A. (1995). Imagery without perception – a case study of anosognosia for cortical blindness. *Neuropsychologia, 33,* 1373-1382.

Goldfield, E. C. (1995). *Emergent forms: origins and early development of human action and perception.* New York: Oxford University Press.

Good, A. & Still, J. (1998). The ontology of mutualism. *Ecological Psychology, 10,* 39-63.

Gottlieb, G. (1998). Normally occurring environmental and behavioral influences on gene activity: From central dogma to probabilistic epigenesis. *Psychological Review, 105,* 792-802.

Gottlieb, G. (1992). *Individual development and evolution.* New York: Oxford University Press.

Gould, S. J. (1991). Exaptation: A crucial tool for an evolutionary psychology. *Journal of Social Issues, 47,* 43-65.

Graf, P. & Schacter, D. L. (1985). Implicit and explicit memory for new associations in normal and amnesic subjects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 501-518.

Gray, R. D. (1992). Death of the gene: Developmental systems strike back. In P. E. Griffiths (Ed.), *Trees of life: Essays in the philosophy of biology* (pp. 165-209). Dordrecht: Kluwer Academic Publishers.

Green, C. D. (1996). Where did the word "cognitive" come from anyway? *Canadian Psychology, 37,* 31-39.

Green, C. D. (1998, April). Are connectionist models theories of cognition? *Psycoloquy, 9.*

Greeno, J. G. & the Middle School Mathematics Through Applications Project Group (1998). The situativity of knowing, learning, and research. *American Psychologist, 53,* 5-26.

Griffiths, P. E. & Gray, R. D. (1994). Developmental systems and evolutionary explanation. *Journal of Philosophy, 91,* 277-304.

Griffiths, P. E. & Gray, R. D. (1997). Replicator II – Judgement Day. *Biology and Philosophy, 12,* 471-492.

Grush, R. (1997). The architecture of representation. *Philosophical Psychology, 10,* 5-23.

Hacking, I. (1983). *Representing and intervening: Introductory topics in the philosophy of natural science.* Cambridge: Cambridge University Press.

Haig, B. D. (1987). Scientific problems and the conduct of research. *Educational Philosophy and Theory, 19,* 22-32.

Haig, B. D. (2000). *An abductive theory of scientific method.* Unpublished manuscript, University of Canterbury.

Haith, M. M. (1998). Who put the cog in infant cognition? Is rich interpretation too costly? *Infant Behavior & Development, 21,* 167-179.

Haken, H. (1978). *Synergetics: An introduction: Nonequilibrium phase transitions and self-organization in physics, chemistry, and biology* (2nd ed.). Berlin: Springer-Verlag.

Haken, H. (1987). Synergetics: An approach to self-organization. In F. E. Yates (Ed.), *Self-organizing systems: The emergence of order* (pp. 417-437). New York: Plenum Press.

Hanson, S. J. & Burr, D. J. (1990). What connectionist models learn: learning and representation in connectionist networks. *Behavioural and Brain Sciences, 13,* 471-518.

Harnad, S. (1990). The symbol grounding problem. *Physica D, 42,* 335-346.

Harré, R. (1976). The constructive role of models. In L. Collins (Ed.). *The use of models in the social sciences* (pp. 16-43). London: Tavistock.

Harré, R. & Gillett, G. (1994). *The discursive mind.* Thousand Oaks: Sage.

Harré, R. & Lamb, R (Eds.). (1983). *The encyclopedic dictionary of psychology.* Cambridge, MA: MIT Press.

Harvey, I., Husbands, P., & Cliff, D. (1993). Issues in evolutionary robotics. In J.-A. Meyer, H. L. Roitblat, & S. W. Wilson (Eds.), *From animals to animats 2: Proceedings of the Second International Conference on Simulation of Adaptive Behavior* (pp. 364-373). Cambridge, MA: MIT Press.

Harvey, I, Husbands, P., & Cliff, D. (1994). Seeing the light: artificial evolution, real vision. In D. Cliff (Ed.), *From animals to animats 3* (pp. 392-401). Cambridge: MIT Press.

Hatano, G. (1982). Cognitive consequences of practice in culture specific procedural skills. *Quarterly Newsletter of the Laboratory of Comparative Human Cognition, 4,* 15-17.

Haugeland, J. (1991). Representational genera. In W. Ramsey, S. P. Stich, & D. E. Rumelhart (Eds.), *Philosophy and connectionist theory* (pp. 61-89). Hillsdale, NJ: Lawrence Erlbaum.

Haugeland, J. (1995). Mind embodied and embedded. In Y.-H. Young & J.-C. Ho (Eds.), *Mind and cognition* (pp. 3-37). Taipei, Taiwan: Academia Sinica.

Hazlehurst, B. & Hutchins, E. (1998). The emergence of propositions from the co-ordination of talk and action in a shared world. *Language and Cognitive Processes, 13,* 373-424.

Heil, J. (1981). Does cognitive psychology rest on a mistake? *Mind, 90,* 321-342.

Held, D. & Thompson, J. B. (Eds.). (1989a). *Social theory of modern societies: Anthony Giddens and his critics.* Cambridge: Cambridge University Press.

Held, D. & Thompson, J. B.(1989b). Editor's Introduction. In D. Held & J. B. Thompson (Eds.), *Social theory of modern societies: Anthony Giddens and his critics* (pp. 1-18). Cambridge: Cambridge University Press.

Hendriks-Jansen, H. (1996). *Catching ourselves in the act: situated activity, interactive emergence, evolution, and human thought.* Cambridge, MA: MIT Press.

Hendriks-Jansen, H. (1996). In praise of interactive emergence, or why explanations don't have to wait for implementations. In M. A. Boden (Ed.), *The philosophy of artificial life* (pp. 282-299). New York: Oxford University Press. (Reprinted from *Artificial life IV*, pp. 70-79, by R. A. Brooks & P. Maes, Eds., 1994, Cambridge MA: MIT Press)

Heyes, C. M. (1998). Theory of mind in nonhuman primates. *Behavioral and Brain Sciences, 21,* 101-148.

Hinton, G. E. (1981). Implementing semantic networks in parallel hardware. In G. E. Hinton & J. A. Anderson (Eds.), *Parallel models of associative memory.* Hillsdale: N. J. Erlbaum.

Hinton, G. E., McClelland, J. L., & Rumelhart, D. E. (1986). Distributed representations. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition, Volume 1: Foundations* (pp. 77-109). Cambridge, MA: MIT Press.

Ho, M.-W. (1986). Heredity as process: Towards a radical reformulation of heredity. *Revista di Biologia/ Biology Forum, 79,* 407-477.

Ho, M.-W. (1993). *The rainbow and the worm: the physics of organisms.* Singapore: World Scientific.

Hodgins J. K. & Brogan, D. C. (1994). Robot herds: group behaviors for systems with significant dynamics. In R. Brooks & P Maes (Eds.), *Artificial Life IV* (pp. 319-324). Cambridge, MA: MIT Press.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction: Processes of inference, learning and discovery.* Cambridge MA: MIT Press.

Hooker, C. A. (1987). *A realistic theory of science.* New York: State University of New York.

Hooker, C. A. (1997). Dynamical systems in development: Review essay of Linda V. Smith & Esther Thelen (Eds.) A dynamics systems approach to development: Applications. *Philosophical Psychology, 10,* 103-112.

Horswill, I. (1992). Characterising adaptation by constraint. In F. Varela & P. Bourgine (Eds.), *Towards a practice of autonomous systems: Proceedings of the first European conference on artificial life* (pp. 58-63). Cambridge MA: MIT Press.

Horswill, I. & Brooks, R. A. (1990). Situated vision in a dynamic world: Chasing objects. *Proceedings of the 7th National Conference on Artificial Intelligence,* 796-800.

Howe, M. L. & Courage, M. L. (1993). On resolving the enigma of infantile amnesia. *Psychological Bulletin, 113,* 305-326.

Howe, M. L. & Courage, M. L. (1997). The emergence and early development of autobiographical memory. *Psychological Review, 104,* 499-523.

Humphreys, G. W. & Riddoch, M. J. (1984). Routes to object constancy: Implications from neurological impairments of object constancy. *Quarterly Journal of Experimental Psychology, 36A,* 385-415.

Husbands, P, Harvey, I, & Cliff, D. (1995). Circle in the round: state space attractors in evolved sighted robots. *Robotics and Autonomous Systems, 15,* 83-106.

Hutchins, E. (1995a). *Cognition in the wild.* Cambridge, MA: MIT Press.

Hutchins, E. (1995b). How a cockpit remembers its speeds. *Cognitive Science, 19,* 265-288.

Hutchins, E. & Hazlehurst, B. (1991). Learning in the cultural process. In C. Langton, C. Taylor, D. Farmer, & S. Rasmussen (Eds.), *Artificial life*: Santa Fe Institute Studies in the Sciences of Complexity, Proc. Vol. X. Redwood City, CA: Addison-Wesley.

Hutchins, E. & Hazlehurst, B. (1995). How to invent a lexicon: The development of shared symbols in interaction. In N. Gilbert & R. Conte (Eds.), *Artificial societies: The computer simulation of social life.* London: UCL Press.

Ingold, T. (1993). Technology, language, intelligence: A reconsideration of basic concepts. In K. R. Gibson & T. Ingold (Eds.), *Tools, language and cognition in human evolution.* (pp. 449-472). Cambridge: Cambridge University Press.

Ingold, T. (1996). Situated action V: The history and evolution of bodily skills. *Ecological Psychology, 8,* 171-182.

Jaeger, H. (1998). Today's dynamical systems are too simple. *Behavioral and Brain Sciences, 21,* 643-644.

Janlert, L. (1987). Modeling change – the frame problem. In Z. W. Pylyshyn (Ed.), *The robot's dilemma: the frame problem in artificial intelligence* (pp. 1-40). Norwood, NJ: Ablex.

Jeannerod, M. (1995). Mental imagery in the motor context. *Neuropsychologia, 33,* 1419-1432.

Johnson, M. (1987). *The body in the mind: The bodily basis of imagination, reason, and meaning.* Chicago: University of Chicago Press.

John-Steiner, V. (1992). Private speech among adults. In R. M. Diaz and L. E. Berk (Eds.), *Private speech: from social interaction to self-regulation* (pp. 285-296). Hillsdale, NJ: Lawrence Erlbaum.

Juarrero, A. (1999). *Dynamics in action: Intentional behavior as a complex system.* Cambridge, MA: MIT Press.

Kadar, E. & Effken, J. (1994). Heideggerian meditations on an alternative ontology for ecological psychology: A response to Turvey's (1992) proposal. *Ecological Psychology, 6,* 297-341.

Karmiloff-Smith, A. (1992). *Beyond modularity: a developmental perspective on cognitive science.* Cambridge, MA: MIT Press.

Karmiloff-Smith, A. (1994). Précis of Beyond Modularity: A developmental perspective on cognitive science. *Behavioral and Brain Sciences, 17,* 693-745.

Kauffman, S. A. (1993). *The origins of order: Self-organization and selection in evolution.* New York: Oxford University Press.

Kauffman, S. A. (1995). *At home in the universe.* London: Penguin.

Keijzer, F. A. (1997). *The generation of behavior: on the function of representation in organism-environment dynamics.* Unpublished doctoral dissertation, Leiden University.

Keijzer, F. A. (1998a). Doing without representations which specify what to do. *Philosophical Psychology, 11,* 269-302.

Keijzer, F. A. (1998b). Some armchair worries about wheeled behavior. In R. Pfeifer, B. Blumberg, J-A. Meyer, & S. W. Wilson (Eds.), *From animals to animats 5* (pp. 13-21). Cambridge, MA: MIT Press.

Keijzer, F. A. & Bem, S. (1996). Behavioral systems interpreted as autonomous agents and as coupled dynamical systems: a criticism. *Philosophical Psychology, 9,* 323-346.

Keijzer, F. A. Bem, S. & van der Heijden, L. (1998). The dynamics of what? *Behavioral and Brain Sciences, 21,* 644-645.

Kelso, J. A. S. (1995). *Dynamic patterns: the self-organization of brain and behaviour.* Cambridge, MA: MIT Press.

Kessen, W. (1981). Early settlements in new cognition. *Cognition, 10,* 167-171.

Kirsh, D. (1991). Today the earwig, tomorrow man? *Artificial Intelligence, 47,* 161-184.

Kirsh, D. (1995). The intelligent use of space. *Artificial Intelligence, 72,* 1-52.

Kirsh, D. & Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cognitive Science, 18,* 513-549.

Knapp, T. J. (1986). The emergence of cognitive psychology in the latter half of the twentieth century. In T. J. Knapp & L. C. Robertson (Eds.), *Approaches to cognition: Contrasts and controversies* (pp. 13-35). Hillsdale, N. J.: Lawrence Erlbaum Associates.

Koch, C. & Davis, J. L. (Eds.). (1994). *Large-scale neuronal theories of the brain.* Cambridge, MA: MIT Press.

Kosslyn, S. M. (1986). Toward a computational neuropsychology of high-level vision. In T. J. Knapp & L. C. Robertson (Eds.), *Approaches to cognition: Contrasts and controversies* (pp. 223-242). Hillsdale, N. J.: Lawrence Erlbaum Associates.

Kosslyn, S. M. (1994). *Image and brain: The resolution of the imagery debate.* Cambridge, MA: MIT Press.

Kosslyn, S. M., Behrmann, M., & Jeannerod, M. (1995). The cognitive neuroscience of mental imagery. *Neuropsychologia, 33,* 1335-1344.

Kosslyn, S. M., Thompson, W. L., Kim, I. J., & Alpert, N. M. (1995, November 30). Topographical representations of mental images in primary visual cortex. *Nature, 378*, 496-498.

Kozulin, A. (1986). Vygotsky in context. In A. Kozulin (Ed.), *Thought and language* (pp. xi-lxi). Cambridge, MA: MIT Press.

Kugler, P. N. & Turvey, M. T. (1987*). Information, natural laws, and the self assembly of rhythmic movement.* Hillsdale, NJ: Erlbaum.

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind.* Chicago: University of Chicago Press.

Lalor, B. J. (1997). It *is* what you think: Intentional potency and anti-individualism. *Philosophical Psychology, 10,* 165-178.

Langton, C. G. (1996). Artificial life. In M. A. Boden (Ed.), *The philosophy of artificial life* (pp. 39-94). New York: Oxford University Press. (Reprinted from *1991 lectures in complex systems*, pp. 189-241, by L. Nadel & D. Stein, Eds., 1992, Reading, MA: Addison-Wesley).

Langton, C. G. (Ed.). (1995). *Artificial life: An overview.* Cambridge, MA: MIT Press.

Laudan, L. (1984). A confutation of convergent realism. In J. Leplin (Ed.), *Scientific realism* (pp. 218-249). Berkeley, CA: University of California Press.

Lauder, G. V. (1992). Biomechanics and evolution: Integrating physical and historical biology in the study of complex systems. *Society for Experimental Biology Seminar Series, 36,* 1-19.

Lave, J. (1988). *Cognition in practice: Mind, mathematics, and culture in everyday life.* Cambridge: Cambridge University Press.

Leslie, A. (1987). Pretence and representation: the origins of "theory of mind." *Psychological Review, 94,* 412-426.

Lewis, D. (1972). Psychophysical and theoretical identification. *Australasian Journal of Philosophy, 50,* 247-258.

Leudar, I. & Costall, A. (1996). Situating action IV: Planning as situated action. *Ecological Psychology, 8(2),* 153-170.

Lloyd, D. E. (1989). *Simple minds.* Cambridge, MA: MIT Press.

Loar, B. (1981). *Mind and meaning.* London: Cambridge University Press.

Lock, A. J. (1980). *The guided reinvention of language.* London: Academic Press.

Lock, A. J. (1991). The role of social interaction in early language development. In N. Krasnegor, D. Rumbaugh, R. Schiefelbusch, & M. Studdert-Kennedy (Eds.). *Biological and behavioral determinants of language development.* (pp. 287-300). Hillsdale, NJ: Erlbaum.

Lock, A. J. (1993). Human language development and object manipulation: Their relevance in ontogeny and its possible relevance for phylogenetic questions. In K. R. Gibson & T. Ingold (Eds.), *Tools, language and cognition in human evolution* (pp. 279-299). Cambridge: Cambridge University Press.

Lock, A. J. (1999). On the recent origin of symbolically-mediated language and its implications for psychological science. In M. C. Corballis and S. E. G. Lea (Eds.), *The descent of mind: psychological perspectives on hominid evolution* (pp. 324-355). Oxford: Oxford University Press.

Lock, A. J. & Colombo, M. (1996). Cognitive abilities in a comparative perspective. In A. Lock and C. R. Peters (Eds.), *Handbook of human symbolic evolution* (pp. 596-643). Oxford: Oxford University Press.

Lock, A. J., Service, V., Brito, A., & Chandler, P. (1989). The social structuring of infant cognition. In G. Bremner & A. Slater (Eds.). *Infant development* (pp. 243-271). London: Erlbaum.

Lock, A. J. & Symes, K. (1996). Social relations, communication, and cognition. In A. Lock and C. R. Peters (Eds.). *Handbook of human symbolic evolution* (pp. 204-235). Oxford: Oxford University Press.

Loftus, E. F. & Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior, 13,* 585-598.

Luria, A. R. (1976). *Cognitive development: its cultural and social foundations.* Cambridge, MA: Harvard University Press.

Lyons, W. (1992). Intentionality and modern philosophical psychology - III. The appeal to teleology. *Philosophical Psychology, 5, (3),* 309-326.

Macdonald, C. & Macdonald, G. (Eds.), (1995). *Connectionism: debates on psychological explanation.* Oxford: Blackwell.

MacKay, D. G. (1992). Constraints on theories of inner speech. In D. Reisberg (Ed.), Auditory imagery (pp. 121-149). Hillsdale, NJ: Lawrence Erlbaum.

MacNeillage, P. F. (1998). Evolution of the mechanism of language output: comparative neurobiology of vocal and manual communication. In J. R. Hurford, M. Studdert-Kennedy, & C. Knight (Eds.), *Approaches to the evolution of language* (pp. 222-241). Cambridge: Cambridge University Press.

Maes, P. (1995). Modeling adaptive autonomous agents. In C. G. Langton (Ed.), *Artificial life: an overview* (pp. 135-162). Cambridge, MA: MIT Press.

Maier, S. F. and Watkins, L. R. (1998). Cytokines for psychologists: Implications of bidirectional immune-to-brain communication for understanding behavior, mood, and cognition. *Psychological Review, 105,* 83-107.

Mandler, J. M. (1992). How to build a baby: II. Conceptual primitives. *Psychological Review, 99,* 587-604.

Margulis, L. & Sagan, D. (1995). *What is life?* London: Weidenfeld and Nicolson.

Markman, A. B. & Dietrich, E. (1998). In defense of representation as mediation. *Psycoloquy.98.9.48.*

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.* New York: W. H. Freeman.

Marr, D. & Nishihara, N. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London: Series B, 200,* 269-294.

Mataric, M. J. (1991). Navigating with a rat brain: A neurobiologically inspired model for robot spatial representation. In J. A. Meyer & Wilson, S. W. (Eds.), *From animals to animats: Proceedings of the first international conference on simulation of adaptive behavior* (pp. 169-175). Cambridge, MA: MIT Press.

Maturana, H. R. & Varela, F. J. (1980). *Autopoiesis and cognition: the realization of the living.* Dordrecht: D. Reidel.

Maturana, H. R. & Varela, F. J. (1988). *The tree of knowledge: The biological roots of human understanding.* Boston: Shambhala.

McCabe, V. & Balzano, G. J. (Eds.), (1986). *Event cognition: An ecological perspective.* Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

McClamrock, R. A. (1995). *Existential cognition.* Chicago: University of Chicago Press.

McClelland, J. L., Rumelhart, D. E. & The PDP Research Group (Eds.). (1986). *Parallel distributing processing, Volume 2: Psychological and biological models.* Cambridge, MA: MIT Press.

Mcguire, P. K., Silbersweigh, D. A., Murray, R. M., David, A. S., Frackowiak, R. S. J., & Frith, C. D. (1996). Functional anatomy of inner speech and auditory verbal imagery. *Psychological Medicine, 26,* 29-38.

Medin, D. L. (1989). Concepts and conceptual structure. *American Psychologist, 44,* 1469-1481.

Meltzoff, A. N. (1988). Infant imitation and memory: nine-month-olds in immediate and deferred tests. *Child Development, 59,* 217-225.

Meltzoff, A. N. (1995). What infant memory tells us about infantile amnesia: Long-term recall and deferred imitation. *Journal of Experimental Child Psychology, 59,* 497-515.

Meltzoff, A. N. & Moore, M. K. (1977, October 7). Imitation of facial and manual gestures by human neonates. *Science, 198,* 75-78.

Meltzoff, A. N. & Moore, M. K. (1998). Object representation, identity, and the paradox of early permanence: Steps toward a new framework. *Infant Behavior & Development, 21,* 201-235.

Merleau-Ponty, M. (1965). *The Structure of Behavior.* (A. L. Fisher, Trans.). London: Methuen.

Merzenich M. M., & Kaas, J. H. (1991). Principles of organization of sensory-perceptual systems in mammals. *Progress in Psychobiology and Physiological Psychology, 9,* 2-42.

Millikan, R. G. (1984). *Language, thought, and other biological categories: New foundations for realism.* Cambridge, MA: MIT Press.

Millikan, R. G. (1993). *White queen psychology and other essays for Alice.* Cambridge, MA: MIT Press.

Mingers, J. (1995). *Self-producing systems: Implications and applications of autopoiesis.* New York: Plenum Press.

Munsat, S. (1990). Keeping representations at bay. *Behavioral and Brain Sciences, 13,* 502-503

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92,* 289-316.

Myers, D. G. (1993). *Social psychology* (4th ed.). New York: McGraw-Hill.

Neisser, U. (1967). *Cognitive psychology.* New York: Appleton-Century-Crofts.

Neisser, U. (1976). *Cognition and reality: Principles and implications of cognitive psychology.* San Francisco: W. H. Freeman.

Nelson, K. (1993). The psychological and social origins of autobiographical memory. *Psychological Science, 4,* 7-14.

Nelson, K. (1996). *Language in cognitive development: the emergence of the mediated mind.* Cambridge: Cambridge University Press.

Newell, A. (1990). *Unified theories of cognition.* Cambridge, MA: Harvard University Press.

Newell, A. & Simon, H. A. (1972). *Human problem solving.* Englewood Cliffs, NJ: Prentice Hall.

Newell, A. & Simon, H. A. (1981). Computer science as empirical inquiry: Symbols and search. In J. Haugeland (Ed.), *Mind design: Philosophy, psychology, artificial intelligence* (pp. 35-66). Cambridge, MA: MIT Press.

Noble, W. & Davidson, I. (1989). On depiction and language. *Current Anthropology, 30*(3), 337-342.

Noble, W. & Davidson, I. (1996). *Human evolution, language and mind: a psychological and archaeological inquiry.* Cambridge: Cambridge University Press.

Norman, D. A. (1993a). Cognition in the head and in the world: An introduction to the special issue on situated action. *Cognitive Science, 17,* 1-6.

Norman, D. A. (1993b). *Things that make us smart: Defending human attributes in the age of the machine.* Reading, MA: Addison-Wesley Publishing Company.

Norman, D. A. & Shallice, T. (1986). Attention to action: Willed and automatic control of behaviour. In R. J. Davidson, G. E. Schwartz, & D. Shapiro (Eds.), *Consciousness and self regulation* (Vol. 4, pp. 1-18). New York: Plenum.

Norton, A. (1995). In R. F. Port & T. van Gelder (Eds.), *Mind as motion: Explorations in the dynamics of cognition.* (pp. 45-68). Cambridge, MA: MIT Press.

Olson, D. R. (1989). Making up your mind. *Canadian Psychology, 30,* 617-627.

Olson, D. R. (1993). The development of representations: the origins of mental life. *Canadian Psychology, 34,* 293-306.

Olson, D. R. (1994). *The world on paper: The conceptual and cognitive implications of writing and reading.* Cambridge: Cambridge University Press.

Olson, D. R. (1996). Towards a psychology of literacy: on the relations between speech and writing. *Cognition, 60,* 83-104.

Oyama, S. (1985). *The ontogeny of information: developmental systems and evolution.* Cambridge: Cambridge University Press.

Oyama, S. (1989). Ontogeny and the central dogma: Do we need the concept of genetic programming in order to have an evolutionary perspective? In M. Gunnar & E. Thelen (Eds.), *Systems and development. Minnesota symposia on child psychology.* (Vol. 22, pp. 1-34). Hillsdale, NJ: Erlbaum.

Oyama, S. (1993). How shall I name thee? The construction of natural selves. *Theory & Psychology, 3,* 471-496.

Palmer, S. E. & Kimchi, R. (1986). The information processing approach to cognition. In T. J. Knapp & L. C. Robertson (Eds.), *Approaches to cognition: Contrasts and controversies* (pp. 37-77). Hillsdale, N. J.: Lawrence Erlbaum Associates.

Parisi, D. (1997). An artificial life approach to language. *Brain and Language, 59,* 121-146.

Pea, R. D. (1993). Practices of distributed intelligence and designs for education. In G. Salomon (Ed.), *Distributed cognition: Psychological and educational implications* (pp. 47-87). Cambridge: Cambridge University Press.

Peirce, C. S. (1931-1958). *Collected papers of Charles Sanders Peirce.* (C. Hartshorne & P. Weiss, Eds.). Cambridge, MA: Harvard University Press. (Original work published, 1839-1914).

Penfield, W. (1958). *The excitable cortex in conscious man.* Liverpool: Liverpool University Press.

Perkins, D. N. (1993). Person-plus: A distributed view of thinking and learning. In G. Salomon (Ed.), *Distributed cognition: Psychological and educational implications* (pp. 88-110). Cambridge: Cambridge University Press.

Perner, J. (1991). *Understanding the representational mind.* Cambridge, MA: MIT Press.

Pfeifer, R. & Scheier, C. (1999). *Understanding intelligence.* Cambridge, MA: MIT Press.

Pfeifer, R. & Verschure, P. (1992a). Beyond rationalism: Symbols patterns and behavior. *Connection Science, 4,* 313-325.

Pfeifer, R. & Verschure, P. (1992b). Distributed adaptive control: A paradigm for designing autonomous agents. In Varela, F. & Bourgine, P. (Eds.), *Toward a practice of autonomous systems: Proceedings of the first European conference on artificial life* (pp. 21-30). Cambridge, MA: MIT Press.

Piaget, J. (1951). *Play, dreams and imitation in childhood.* London: Routledge & Kegan Paul.

Piaget, J. (1970). Piaget's theory. In P. H. Mussen (Ed.), *Carmichaels' handbook of child development* (vol. 1, pp. 703-732). New York: Wiley.

Pinker, S. (1994). *The language instinct: How the mind creates language.* New York: William Morrow and Company, Inc.

Pinker, S. (1997). *How the mind works.* New York: W. W. Norton.

Pinker, S. and Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition, 28,* 73-193.

Place, U. T. (1956). Is consciousness a brain process? *British Journal of Psychology, 47,* 44-50.

Pollack, J. (1990). Recursive distributed representations. *Artificial Intelligence, 46,* 77-105.

Pons, T. P., Garraghty, P. T., Ommaya, A. K., Kaas, J. H., Taub, E., & Mishkin, M. (1991, June 28). Massive cortical reorganization after sensory deafferentation in adult macaques [see comments]. *Science, 252,* 1857-1860.

Port, R. F. and Van Gelder, T. (Eds.). (1995). *Mind as motion: Explorations in the dynamics of cognition.* (pp. 195-225). Cambridge, MA: MIT Press.

Premack, D. and Woodruff, G. (1978). Does a chimpanzee have a theory of mind? *Behavioral and Brain Sciences, 1,* 512-526.

Prigogine, I. & Stengers, I. (1985). *Order out of chaos: Man's new dialogue with nature.* London: Flamingo.

Putnam, H. (1975). *Mind, language and reality: Philosophical papers, volume 2.* Cambridge: Cambridge University Press.

Putnam, H. (1988). *Representation and reality.* Cambridge, MA: MIT Press.

Pylyshyn, Z. (1991). The role of cognitive architecture in theories of cognition. In K. VanLehn (Ed.), *Architectures for intelligence* (pp. 189-223). Hillsdale, NJ: Erlbaum.

Pylyshyn, Z. (Ed.). (1987). *The robot's dilemma: The frame problem in artificial intelligence.* Norwood, NJ: Ablex.

Ramachandran, V. S. (1988, August). Perception of shape from shading. *Scientific American, 256,* 76-83.

Real, L. A. (1991, August 30). Animal choice behavior and the evaluation of cognitive architecture. *Science, 253,* 980-986.

Redington, M. & Chater, N. (1998). Connectionist and statistical approaches to language acquisition: A distributional perspective. *Language and Cognitive Processes, 13,* 129-191.

Reed, E. S. (1993). The intention to use a specific affordance: A conceptual framework for psychology. In R. H. Wozniak & K. W. Fischer (Eds.), *Development in context: Acting and thinking in specific environments* (pp. 45-76). Hillsdale, New Jersey: Lawrence Erlbaum Associates, Publishers.

Reed, E. S. (1996). *Encountering the world: toward an ecological psychology.* New York: Oxford University Press.

Reeke, G. N. & Sporns, O. (1990). Selectionist models of perceptual and motor systems and implications for functionalist theories of brain function. *Physica D, 42,* 347-364.

Reisberg, D. (Ed.). (1992). *Auditory imagery.* Hillsdale, NJ: Lawrence Erlbaum.

Rensink, R. A., O'Regan, J. K., & Clark, J. J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science, 8,* 368-373.

Resnick, M. (1994). *Turtles, termites, and traffic jams: Explorations in massively parallel microworlds.* Cambridge, MA: MIT Press.

Rey, G. (1997). *Contemporary philosophy of mind: A contentiously classical approach.* Cambridge, MA: Blackwell.

Reynolds, C. W. (1987). Flocks, herds, and schools: A distributed behavioral model. *Computer Graphics, 21,* 25-34.

Roediger, H. L. & Blaxton, T. A. (1987). Retrieval modes produce dissociations in memory for surface information. In D. S. Gorfein & R. R. Hoffman (Eds.), *Memory and cognitive processes: The Ebbinghaus centennial conference* (pp. 349-379). Hillsdale, NJ: Lawrence Erlbaum.

Rogoff, B. (1990). *Apprenticeship in thinking: Cognitive development in social context.* New York: Oxford University Press.

Rosch, E. & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7,* 573-605.

Rowlands, M. (1999). *The body in mind: Understanding cognitive processes.* Cambridge: Cambridge University Press.

Rumelhart, D. E. (1989). The architecture of mind: A connectionist approach. In M. I. Posner (Ed.), *Foundations of cognitive science* (pp. 133-159). Cambridge, MA: MIT Press.

Rumelhart, D. E., Smolensky, P., McClelland, J. L., & Hinton, G. E. (1986). Schemata and sequential thought processes in PDP models. In J. L. McClelland, D. E. Rumelhart, & The PDP Research Group (Eds.), *Parallel distributing processing, Volume 2: Psychological and biological models* (pp. 7-57). Cambridge, MA: MIT Press.

Runeson, S. (1977). On the possibility of " smart" perceptual mechanisms. *Scandinavian Journal of Psychology, 18,* 172-179.

Russell, J. (1996). Development and evolution of the early symbolic function: the role of working memory. In P. Mellars & K. Gibson (Eds.), *Modelling the early human mind* (pp. 159-170). Cambridge: The McDonald Institute for Archaeological Research.

Rutkowska, J. (1994). Emergent functionality in human infants. In D. Cliff, P. Husbands, J-A. Meyer, & S. W. Wilson (Eds.), *From animals to animats 3* (pp. 179-188). Cambridge, MA: MIT Press.

Ryle, G. (1949). *The concept of mind.* London: Hutchinson.

Salomon, G. (Ed.), (1993a). *Distributed cognition: Psychological and educational implications.* Cambridge: Cambridge University Press.

Salomon, G. (1993b). No distribution without individual's cognition: A dynamic interactional view. In G. Salomon (Ed.), *Distributed cognition: Psychological and educational implications* (pp. 111-138). Cambridge: Cambridge University Press.

Saltzman, E. L. (1995). Dynamics and coordinate systems in skilled sensorimotor activity. In R. F. Port & T. van Gelder (Eds.), *Mind as motion: Explorations in the dynamics of cognition.* (pp. 149-173). Cambridge, MA: MIT Press.

Sarter, M., Berntson, G. G., & Cacioppo, J. T. (1996). Brain imaging and cognitive neuroscience. *American Psychologist, 51,* 13-21.

Schacter, D. L. (1987). Implicit memory: History and current status. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 501-518.

Schechtman, M. (1997). The brain/body problem. *Philosophical Psychology, 10,* 149-165.

Scheerer, E. (1996). Orality, literacy, and cognitive modeling. In B. M. Velichkovsky & D. M. Rumbaugh (Eds.), *Communicating meaning: the evolution and development of language.* (pp. 211-256). Mahwah, NJ: Lawrence Erlbaum.

Schrodinger, E. (1945). *What is life? The physical aspect of the living cell.* Cambridge: Cambridge University Press.

Scribner, S. (1986). Thinking in action: Some characteristics of practical thought. In R. J. Sternberg & R. K. Wagner (Eds.), *Practical intelligence: Nature and origins of competence in the everyday world* (pp. 13-30). Cambridge: Cambridge University Press.

Scribner, S. & Cole, M. (1981). *The psychology of literacy.* Cambridge, MA: Harvard University Press.

Searle, J. R. (1992). *The rediscovery of the mind.* Cambridge, MA: MIT Press.

Segalowitz, S. J. & Bernstein, D. (1997). Neural networks and neuroscience. What are connectionist simulations good for? In D. M. Johnson (Ed.), *The future of the cognitive revolution* (pp. 209-216). New York: Oxford University Press.

Seidenberg, M. S. (1997, March 14). Language acquisition and use: Learning and applying probabilistic constraints. *Science, 275,* 1599-1603.

Sejnowski, T. J. & Rosenberg, C. R. (1987). Parallel networks that learn to pronounce English text. *Complex Systems, 1,* 145-168.

Shanon, B. (1991). Representations: senses and reasons. *Philosophical Psychology, 4,* 355-374.

Shanon, B. (1992). Are connectionist models cognitive? *Philosophical Psychology, 5,* 235-255.

Sherry, D. F. & Schacter, D. L. (1987). The evolution of multiple memory systems. *Psychological Review, 94,* 439-454.

Simon, H. A. (1981). *The sciences of the artificial* (2nd ed.). Cambridge, MA: MIT Press.

Simpson, J. A. & Weiner, E. S. C. (Eds.). (1989). *The Oxford English Dictionary.* Oxford: Clarendon Press.

Skinner, B. F. (1957). *Verbal behavior.* New York: Appleton-Century-Crofts.

Skinner, B. F. (1989). The origins of cognitive thought. *American Psychologist, 44,* 13-18.

Slocum, A.C., Downey, D.C. and Beer, R.D. (2000). Further experiments in minimally cognitive behavior: From perceiving affordances to selective attention. To appear in *From Animals to*

Animats 6: Proceedings of the Sixth International Conference on Simulation of Adaptive Behavior.

Smart, J. J. C. (1962). Sensations and brain processes. In V. C. Chappell (Ed.), *The philosophy of mind* (pp. 160-172). Englewood Cliffs, NJ: Prentice-Hall.

Smith, B. C. (1996). *On the Origin of Objects.* Cambridge: MIT Press.

Smith, J. D., Reisberg, D., & Wilson, M. (1992). Subvocalization and auditory imagery: Interactions between the inner ear and the inner voice. In D. Reisberg (Ed.), *Auditory imagery* (pp. 95-119). Hillsdale, NJ: Lawrence Erlbaum.

Smith, L. B., Thelen, E., Titzer R., & McLin. D. (1999). Knowing in the context of acting: The task dynamics of the A-not-B error. *Psychological-Review, 106,* 235-260.

Smith, S. M., Brown, H. O., Tolman, J. E. P. & Goodman, L. S. (1947). The lack of cerebral effects of d-Tubercurarine. *Anesthesiology, 8,* 1-14.

Smithers, T. (1992). Taking eliminative materialism seriously: A methodology for autonomous systems research. In F. Varela & P. Bourgine (Eds.), *Towards a practice of autonomous systems: Proceedings of the first European conference on artificial life* (pp. 31-40). Cambridge MA: MIT Press.

Smithers, T. (1995). Are autonomous agents information processing systems? In L. Steels & R. A. Brooks (Eds.), *The artificial life route to artificial intelligence: Building situated embodied agents* (pp. 123-162). Hillsdale, NJ: Erlbaum.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences, 11,* 1-74.

Smolensky, P. (1991). Connectionism, constituency and the language of thought. In B. Loewer & G. Rey (Eds.), *Meaning in mind: Fodor and his critics.* Oxford: Basil Blackwell.

Smolensky, P. (1995). Reply: Constituent structure and explanation in an integrated connectionist/symbolic cognitive architecture. In C. Macdonald, & G. Macdonald (Eds.), *Connectionism: debates on psychological explanation.* (pp. 223-290). Oxford: Blackwell.

Smythe, W. E. (1989). The case for cognitive conservatism: A critique of Dan Lloyd's approach to mental representation. *Behaviorism, 17, (1),* 63-73.

Snowden, P. F. (1988). Review of representation and reality by D. Papineau, *Mind, 97,* 629-632.

Sokolov, A. N. (1972). *Inner speech and thought* (G. T. Onischenko, Trans.). New York: Plenum. (Original work published 1968).

Spelke, E. S. (1990). Principles of object perception. *Cognitive Science, 14,* 29-56.

Spelke, E. S. (1998). Nativism, empiricism, and the origins of knowledge. *Infant Behavior & Development, 21,* 181-200.

St John, M. F. & McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence, 46,* 159-257.

Steels, L. (1995). The artificial life roots of artificial intelligence. In C. G. Langton (Ed.),*Artificial life: an overview* (pp. 75-110). Cambridge, MA: MIT Press.

Steels, L. (1998). Synthesizing the origins of language and meaning using coevolution, self-organization and level formation. In J. R. Hurford, M. Studdert-Kennedy, & C. Knight (Eds.), *Approaches to the evolution of language* (pp. 384-404). Cambridge: Cambridge University Press.

Sterelny, K. (1990). *The representational theory of mind.* Oxford: Basil Blackwell.

Sterelny, K. & Kitcher, P. (1988). The return of the gene. *Journal of Philosophy, 85,* 339-361.

Sterelny, K., Smith, K. C. & Dickison, M. (1996). The extended replicator. *Biology and Philosophy, 11,* 377-403.

Sternberg, R. J. (Ed.). (1998). *The nature of cognition.* Cambridge, MA: MIT Press.

Stewart, I. & Cohen, J. (1997). *Figments of reality: the evolution of the curious mind.* Cambridge: Cambridge University Press.

Stewart, I. (1997). *Does God play dice? The new mathematics of chaos* (Rev. ed.). London: Penguin.

Stewart, I. (1998). *Life's other secret: The new mathematics of the living world.* London: Penguin

Stewart, J. (1995).Cognition = life: Implications for higher-level cognition,*Behavioural Processes, 35,* 311-326.

Stich, S. (1983). *From folk psychology to cognitive science.* Cambridge, MA. MIT Press.

Suchman, L. A. (1987). *Plans and situated actions: the problem of human-machine communication.* Cambridge, MA: Cambridge University Press.

Suchman, L. A. (1993). Response to Vera and Simon's situated action: A symbolic interpretation, *Cognitive Science, 17,* 71-86.

Suddendorf, T. (1999). The rise of the metamind. In M .C. Corballis & S. E. G. Lea (Eds.). *The descent of the mind: psychological perspectives on hominid evolution.* (pp. 218-260). Oxford: Oxford University Press.

Sutton, J. (1998). Being there: putting philosopher, researcher and student together again. Review of "Being there: putting brain, body, and world together again." *Metascience, 7,* 90-95.

Swenson, R. & Turvey, M. T. (1991). Thermodynamic reasons for perception-action cycles, *Ecological-Psychology, 3,* 317-348.

Taga, G. (1994). Emergence of bipedal locomotion through entrainment among the neuro-musculo-skeletal system and the environment. *Physica D, 75,* 190-208.

Taga, G., Yamaguchi, Y., & Shimizu, H. (1991). Self-organized control of bipedal locomotion by neural oscillators in unpredictable environment. *Biological Cybernetics, 65,* 147-159.

Tarr, M. J. & Pinker, S. (1990). When does human object recognition use a viewer-centered reference frame? *Psychological Science, 1,* 253-256.

Thelen, E. (1986). Treadmill-elicited stepping in seven-month-old infants, *Child-Development, 57,* 1498-1506.

Thelen, E. (1991). Motor aspects of emergent speech: A dynamic approach. In N. Krasnegor, D. Rumbaugh, R. Schiefelbusch, & M. Studdert-Kennedy (Eds.), *Biological and behavioral determinants of language development.* (pp. 339-362). Hillsdale, NJ: Erlbaum.

Thelen, E. (1995). Time-scale dynamics and the development of an embodied cognition. In R. F. Port & T. van Gelder (Eds.), *Mind as motion: Explorations in the dynamics of cognition* (pp. 69-100). Cambridge, MA: MIT Press.

Thelen, E., Corbetta, D., Kamm, K., Spencer, J. P., Schneider, K. & Zernicke, R. (1993). The transition to reaching: Mapping intention and intrinsic dynamics. *Child-Development, 64,* 1058-1098.

Thelen, E. & Fisher, D. M. (1982). Newborn stepping: An explanation for a "disappearing reflex." *Developmental Psychology, 18,* 760-775.

Thelen, E., Fisher, D. M., Ridley-Johnson, R., & Griffin, N. J. (1982). Effects of body build and arousal on newborn infant stepping, *Developmental Psychobiology, 15,* 447-453.

Thelen, E. & Smith, L. B. (1994). *A dynamic systems approach to the development of cognition and action.* Cambridge, MA: MIT Press.

Thomas, N. J. T. (1999). Are theories of imagery theories of imagination? An active perception approach conscious mental content. *Cognitive Science, 23,* 207-245.

Tomasello, M. (1996). The cultural roots of language. In B. M. Velichkovsky & D. M. Rumbaugh (Eds.), *Communicating meaning: the evolution and development of language* (pp. 275-307). Mahwah, NJ: Lawrence Erlbaum.

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of memory* (pp. 382-403). New York: Academic.

Turvey, M. T. (1990). Coordination. *American Psychologist, 45*(8), 938-953.

Turvey, M. T., Shaw, R. E., Reed, E. S., & Mace, W. M. (1981). Ecological laws of perceiving and acting: In reply to Fodor and Pylyshyn. *Cognition, 9,* 237-304.

Ullman, S. (1980) Against direct perception. *Behavioral and Brains Sciences, 3,* 373-416.

Valsiner, J. (1991). Construction of the mental: From the 'cognitive revolution' to the study of development. *Theory & Psychology, 1,* 477-494.

Valsiner, J. (1997). *Culture and the development of children's action: a theory of human development* (2nd ed.). New York: John Wiley & Sons.

van Gelder, T. (1995). What might cognition be, if not computation? *The Journal of Philosophy, 91,* 345-381.

van Gelder, T. (1998). The dynamical hypothesis in cognitive science. *Behavioral and Brain Sciences, 21,* 615-665.

van Gelder, T. & Port, R. F. (1995). Its about time: An overview of the dynamical approach to cognition. In R. F. Port & T. van Gelder (Eds.), *Mind as motion: Explorations in the dynamics of cognition* (pp. 1-39). Cambridge, MA: MIT Press.

Varela, F. J., Thompson, E., & Rosch, E. (1991). *The embodied mind: Cognitive science and human experience.* Cambridge, MA: MIT Press.

Vera, A. H. & Simon, H. A. (1993). Situated action: A symbolic interpretation. *Cognitive Science, 17,* 7-48.

Verschure, P. F. M. J. (1992). Taking connectionism seriously: The vague promise of subsymbolism and an alternative. In *Proceedings of the 14th Annual Conference of the Cognitive Science Society* (pp. 653-658). Hillsdale, NJ: Erlbaum.

Vicente, K. J. & Burns, C. M. (1996). Evidence for direct perception from cognition in the wild, *Ecological-Psychology, 8,* 269-280.

Von Eckardt, B. (1993). *What is cognitive science?* Cambridge, MA: MIT Press.

Von Uexküll, J. (1957). A stroll through the worlds of animals and men. In C. H. Schiller & K. S. Lashley (Eds.), *Instinctive Behaviour: The development of a modern concept* (pp. 5-82). New York: International University Press. (Original work published 1934)

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* (M. Cole, V. John-Steiner, S. Scribner, & E. Souberman, Trans.). Cambridge, MA: Harvard University Press. (Original work published 1930, 1933, 1935).

Vygotsky, L. S. (1986). *Thought and language* (A. Kozulin, Trans.). Cambridge, MA: MIT Press. (Original work published 1934).

Vygotsky, L. S. and Luria, A.R. (1993) *Studies on the history of behavior: Ape, primitive, and child* (V. I. Golod & J. E. Knox, Trans.). Hillsdale, NJ: Erlbaum.

Wertsch, J. V. (1985). *Vygotsky and the social formation of mind.* Cambridge, MA: Harvard University Press.

Wertsch, J. V. (1991). *Voices of the mind: a sociocultural approach to mediated action.* Cambridge, MA: Harvard University Press.

Wessells, N. K. & Hopson, J. L. (1988). *Biology.* New York : Random House.

Wexler, M., Kosslyn, S., & Berthoz, A. (1998). Motor processes in mental rotation. *Cognition, 68,* 77-94.

Wheeler, M. (1996). From robots to Rothko: the bringing forth of worlds. In M. A. Boden (Ed.), *The philosophy of artificial life* (pp. 209- 236). Oxford: Oxford University Press.

Wheeler, M. (1998). An appeal for liberalism, or why van Gelder's notion of a dynamical system is too narrow for cognitive science. *Behavioral and Brain Sciences, 21,* 653-654.

Whitcombe, E. (1996). The anatomical foundations of cognition: suggestions for a reinterpretation. In P. Mellars & K. Gibson (Eds.), *Modelling the early human mind* (pp. 81-87). Cambridge: The McDonald Institute for Archaeological Research.

Whitehead, S. D. & Ballard, D. H. (1990). Active perception and reinforcement learning. *Neural Computation, 2,* 409-419.

Wimmer, H. & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13,* 103-128.

Winograd, T. & Flores, F. (1986). *Understanding computers and cognition: A new foundation for design.* Reading, MA: Addison Wesley.

Wood, C. C. (1976). Discriminability, response bias, and phoneme categories on discrimination of voice onset time. *Journal of the Acoustical Society of America, 60,* 1381-1389.

Wood, D. J., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry, 17,* 89-100.

Wozniak, R. H. (1976). Speech-for-self as a multiply reafferent human action system. In K. F. Riegel and J. A. Meacham (Eds.). *The developing individual in a changing world: historical and cultural views* (Vol. 1, pp. 151-160). The Hague: Mouton.

Wynn, K. (1992). Children's acquisition of the number words and counting system. *Cognitive Psychology, 24,* 220-251.

Zatorre, R. J., Halpern, A. R., Perry, D. W., Meyer, E., & Evans, A. C. (1996). Hearing in the mind's ear. *Journal of Cognitive Neuroscience, 8,* 29-46.

Zelinsky, G. J., Rao, R. P. N., Hayhoe, M. M., & Ballard, D. H. (1997). Eye movements reveal the spatiotemporal dynamics of visual search. *Psychological Science, 8,* 448-453.